Functional Characterisation of Putative Cis-Regulatory Risk Loci for Breast Cancer.

By

Nordiana Rosli

Thesis for the degree of

Erasmus Mundus Master in Quality in Analytical Laboratories

University of Algarve, Portugal

2015

Functional Characterisation of Putative Cis-Regulatory Risk Loci for Breast Cancer.

By
Nordiana Rosli

Thesis for the degree of
Erasmus Mundus Master in Quality in Analytical Laboratories

Supervisor
Professor Doctor Ana Teresa Maia, PhD

Co-Supervisor
Doctor Joana Xavier, PhD

Department of Biomedical Sciences and Medicine
University of Algarve, Portugal
2015

Acknowledgements

First of all, I am grateful to God for the good health and well-being that were necessary in order to complete this thesis.

I wish to express my sincere thanks to the European Commission, for providing the scholarship for the EMQAL programme entrance.

I place on record, my sincere thank you to the EMQAL management team, educators and student members for the continuous encouragement.

I wish to express my deepest gratitude to Professor Ana Teresa Maia, Doctor Joana Xavier as well as fellow laboratory members, Catia, Bernardo and Marlene, for their genuine apprehension, encouragement, patient, guidance and whose expertise and knowledge were generously shared.

To my beloved family and friends, especially my Mak, Angah and Adek for their unconditional love, unceasing encouragement, support and attention.

In loving memories of Ayah. Because of you, I took this journey. Thank you and you will always be remembered.

Finally, I place on record my sense of gratitude to one and all, who directly and indirectly, have lent their hand in this venture.

Abstract

At present, 94 breast cancer susceptibility loci have been discovered from genome-wide association studies (GWAS). The next step is to identify the causal risk variant, the target gene and to understand the underlying disease mechanism. Studies revealed that most of the variants discovered by GWAS are cis-acting regulatory. Cis-acting regulatory variants can be identified most efficiently by differential allelic expression (DAE) analysis. A DAE genome-wide mapping was done in normal breast tissue, which was cross-compared with GWAS breast cancer data. 19 loci associated with risk and with evidence of cis-regulation were identified, including the 5q14.2 locus that has one SNP associated with risk - rs7707921, and five SNPs displaying DAE across three genes: *ATG10, RPS23* and *ATP6AP1L*.

The aim of this thesis is to set out to map the regulatory variants responsible for the DAE signals in the 5q14.2 locus and to determine which one(s) is (are) associated with risk for breast cancer.

We performed *in silico* analysis using data obtained through publically accessible databases, to identify candidate regulatory SNPs (rSNPs) that could be responsible for the DAE and determine if they may be associated with risk to breast cancer. Experimental *in vitro* analysis by EMSA and analysis of available ChIP-seq data was also conducted in order to investigate possible interactions between candidate rSNPs and transcription factors (TFs).

In this study, three SNPs rs226198, rs6880209 and rs17247678 were identified as potential cis-acting regulators of *ATG10, RPS23* and *ATP6AP1L*. Henceforth, we propose a risk model based on our findings: Binding of c-Myc and POL2 to the common allele of rs226198 and rs6880209 lead to over expression of *RPS23* and under expression of *ATG10*, respectively, whereas, binding of STAT3 and c-FOS to rs17247678 lead to under expression of *ATP6AP1L*, increasing the risk for breast cancer.

Keywords: breast cancer; genetic risk; cis-acting regulatory variants; differential allelic expression

Table of Contents

List of Figures

List of Tables

List of Annex

List of Abbreviation

3C	Chromatin Conformation Capture
4C	Circularised Chromatin Conformation Capture
5C	Carbon-Copy Chromatin Conformation Capture
CCLE	Cancer Cell Line Encyclopedia
cDNA	Complementary Deoxyribonucleic Acid
ChIP-seq	Chromatin Immunoprecipitation Sequencing
DAE	Differential Allelic Expression
DHS	DNase I Hypersensitivity
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
EMSA	Electrophoretic Mobility Shift Assay
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression Quantitative Trait Loci
ER-	Oestrogen Receptor Negative
ER+	Oestrogen Receptor Positive
FRET	Fluorescent Resonance Energy Transfer
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Study
HMEC	Human Mammary Epithelial Cells
HMF	Human Mammary Fibroblasts
IGV	Integrative Genomics Viewer
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
PCR	Polymerase Chain Reaction
PWM	Position Weight Matrix
qPCR	Real-time Polymerase Chain Reaction
RNA	Ribonucleic Acid

RPMI	Roswell Park Memorial Institute
rSNP	Regulatory Single Nucleotide Polymorphism
SNP	Single Nucleotide Polymorphism
TF	Transcription Factor
tSNP	Transcribed Single Nucleotide Polymorphism
WHO	World Health Organisation

Chapter 1: Introduction

1.1 Cancer

In the human body, new cells will grow and divide in order to replace old or damaged cells that will eventually die. However, when these new cells start to grow out of control, it can lead to serious illness or death. This abnormal cellular growth is called cancer. It can start anywhere within the human body then invade and spread to other parts of the body.

In 2000, Hanahan et. al. suggested six key elements for cancer development. These key elements, named "The Hallmarks of Cancer", include activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, resisting cell death, sustaining proliferative signalling and evading growth suppressor[1,2]. Additional four hallmarks were added in 2011, these are avoiding immune destruction, tumour-promoting inflammation, deregulating cellular energetics and genome instability and mutation (Figure 1)[2].

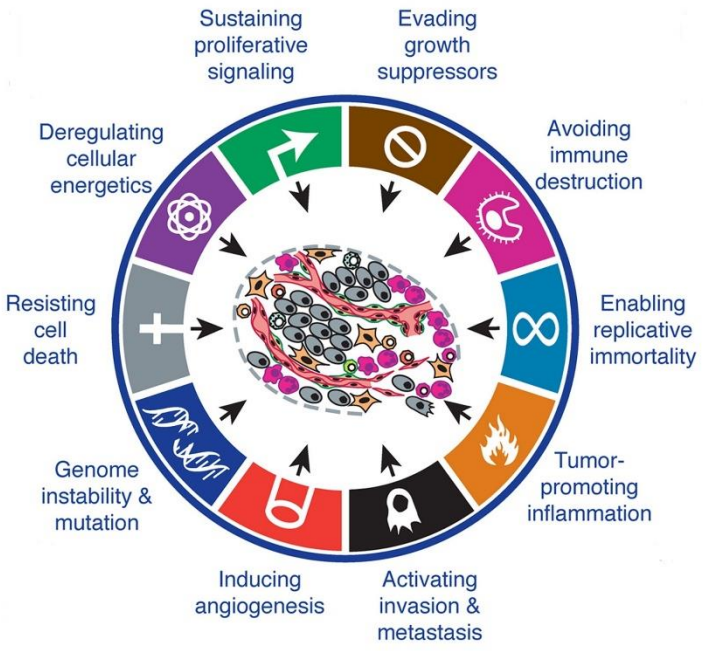


Figure 1 : Hallmarks of Cancer. Modified from Hanahan et. al., 2011.

Genomic instability is a major leading force for carcinogenesis[3]. In hereditary cancers, the establishment of genomic instability could be the launching pad, in which it facilitates the beginning of all the other hallmarks. Meanwhile for non-hereditary cancers, the deregulation of proliferation regulating genes could be the start-up reason. Subsequently, this leads to DNA damage and DNA replication stress, which then points to genomic instability and selective pressure for tumour suppressor p53 inactivation. Failure of the p53 function allows cell evasion from cell death and with genomic instability, it provides a fertile ground for additional mutations that lead to the start of the remaining hallmarks[4].

Cancer contributes amongst the leading number of morbidity and mortality worldwide, having 14.1 million new cases and 8.2 million deaths reported in GLOBOCAN 2012[5]. In relation to these numbers, it is estimated that 24.4% of cancer incidence and 21.5% of cancer mortality occurs in Europe (Figure 2).

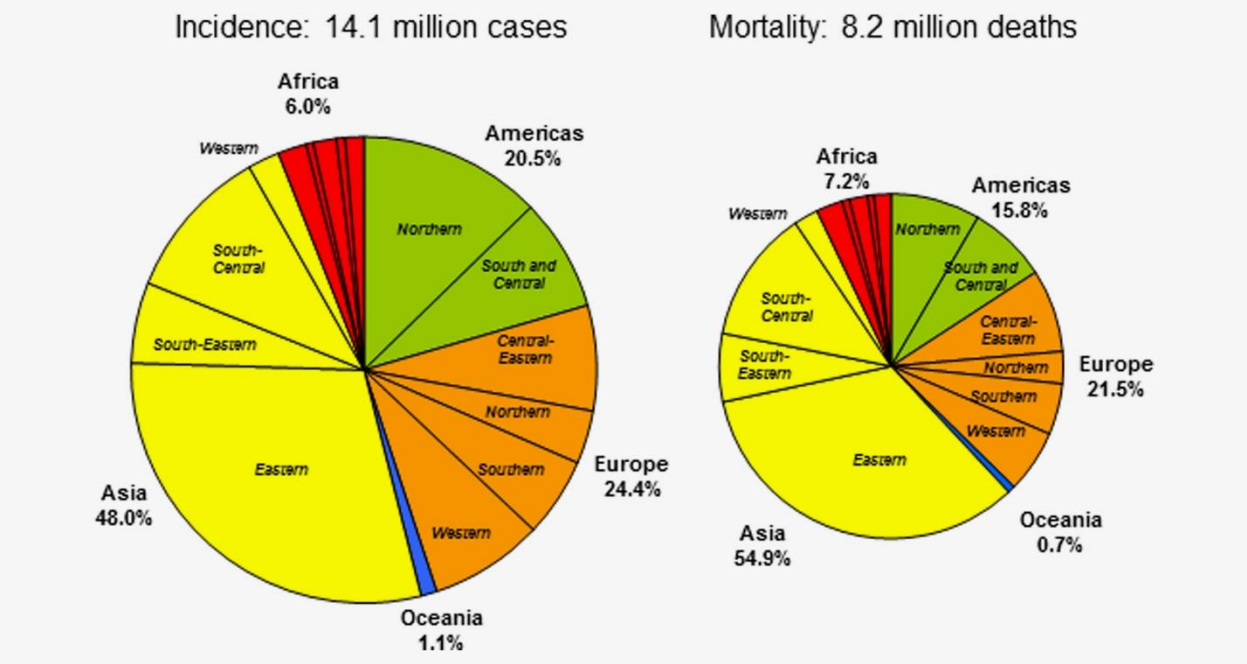


Figure 2 : Estimated numbers of incidence and mortality cases by world regions, 2012. Ferlay et. al., 2015.

In 2015, World Health Organisation (WHO) estimates that the number of new cases will rise by about 70% in the next 20 years. The most common cancer that occurs among men are

lung, prostate, colorectum, stomach and liver cancer. Meanwhile in women, the most common cancer occurrence are breast, colorectum, lung, cervix and stomach cancer[5].

1.2 Breast Cancer

1.2.1 Epidemiology

As mentioned previously, breast cancer occurs most frequently amongst women, with 25% overall of cancer cases in women worldwide. According to GLOBOCAN 2012, there were about 1.67 million cases of breast cancer and this contributed to about 11.9% of the total cancer cases worldwide[5]. Subsequently, this puts breast cancer as the 5th most deadly cancer with 522,000 deaths per year, making up 6.4% of overall cancer deaths.

In Europe, it is estimated that 464,200 new breast cancer cases arise per year, meanwhile 131,300 cases leads to death. Incidence rates differ by 10.1% between the more developed regions with the less developed regions. However, mortality rates between these classes of regions differ by 39.7%. The reason for this difference is thought to be due to patients having better survival rates due to better access of medical treatment in the more developed regions[5].

1.2.2 Risk Factors

There are many factors that can contribute to the development of breast cancer. Non-inherited risk factors such as age, weight, usage of hormonal contraception, hormone replacement treatment, dietary and lifestyle, could lead to the development of breast cancer in women[6,7]. Meanwhile women with family history of breast cancer have higher risk of developing the disease, compared to women who do not have any family history[8,9]. This genetic risk is further evident by the high concordance of breast cancer cases among monozygotic twins, compared to dizygotic twins or siblings[10].

1.2.3 Genetic Susceptibility

Genetic susceptibility is a complex matter due to variation at many different loci. The susceptibility alleles at these loci have different frequencies and impose different cancer risks[10]. These genetic variants can be categorised into three groups, high penetrance alleles, moderate penetrance alleles and low penetrance alleles (Figure 3). Penetrance refers to the relative risk of developing cancer, conferred by an allele with a given population frequency.

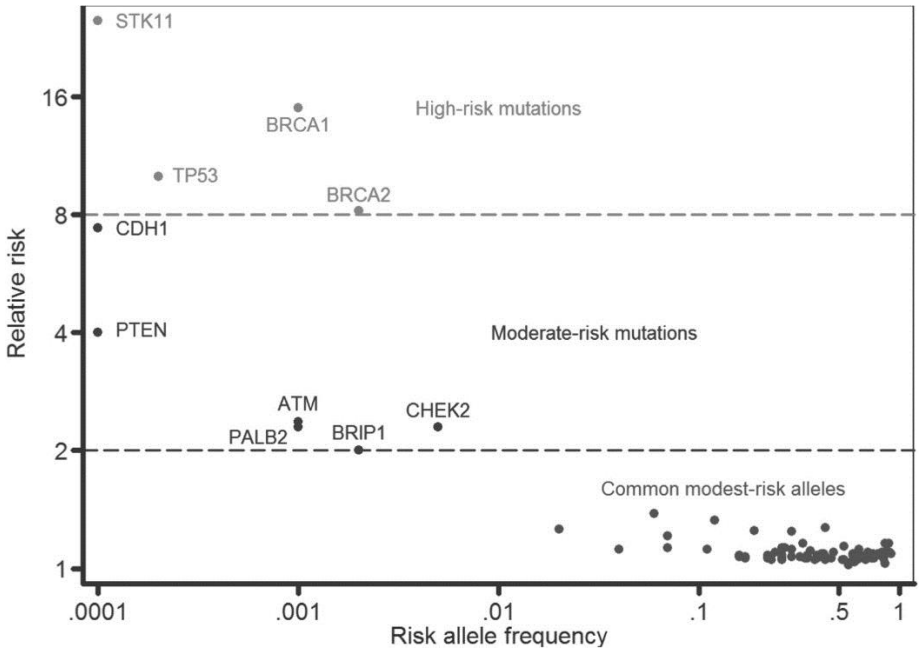


Figure 3 : Characterisation of breast cancer genetic susceptibility. In y-axis, it refers to the relative risk meanwhile in x-axis it refers to the risk allele frequency. Ghoussaini et. al., 2013.

i. High Penetrance Alleles

High penetrance alleles confer a high risk of developing cancer, with > 50% of lifetime risk. However, these alleles are rare in the population. *BRCA1* and *BRCA2* germ line mutations are amongst the few high penetrance alleles that were discovered back in 1990's. These tumour suppressor genes are involved in DNA repair and confer a 10 - 30 fold increase in risk of developing cancer in carriers, compare to non-carriers[11]. Few other high penetrance alleles have been identified, such as, mutations in *TP53* and *STK11/LKB1*[10].

ii. Moderate Penetrance Alleles

Moderate penetrance alleles confer a moderate risk of developing cancer, with ≥ 20% of lifetime risk. Alleles that have been identified and classified into this group include mutations and polymorphisms in genes *ATM, CHEK2, PALB2, BRIP1, PTEN and CDH1*. These alleles confer a 2 - 8 fold increase in risk to breast cancer[10].

iii. Low Penetrance Alleles

Low penetrance alleles, were more recently identified through genome-wide associations studies, and confer a low risk of developing cancer, normally with 10 - 20% of lifetime risk. However, these alleles are common in the population with minor allele frequencies of > 5%, and are associated with a modest to low increase in breast cancer risk (< 1.5 fold)[10].

1.3 Genome-Wide Association Study

Common low penetrance alleles are generally identified by case-control association studies such as genome-wide association studies (GWAS). GWAS functions by examining and comparing common DNA sequence variations in a large set of unrelated cases and controls in order to associate differences in allele frequencies at a polymorphic marker, with the trait of interest. However, this study model must be well-matched to avoid population stratification, as unmatched cases and controls study could lead to bias results, due to differences in race or ethnicity of those with and without the disease of interest.

The genetic variations in GWASes are generally focused on single nucleotide polymorphisms (SNPs) and often these GWAS associated SNPs are located outside of genes in non-coding regions. So far, GWASes were able to identify hundreds of cancer susceptibility loci and currently there are 94 loci associated with breast cancer, which altogether explain ~ 16% of familial risk of this disease[12].

Even though GWASes have enabled the association between genetic variations and risk for diseases, they were unable to precisely identify the association exact target. Therefore the biggest post-GWAS challenges are to (1) identify the causal variants and (2) the genes that are linked to these associations. Afterwards, further studies need to be done in order to understand the functional consequences of these causal variants and to correctly define the mechanism by which these genetic variations and affiliated genes act.

1.4 Single Nucleotide Polymorphism

There are millions of genetic variations with a minor allele frequency (MAF) of > 1% in the population. MAF (minor allele frequency) refers to the frequency at which the least common allele occurs in the population. These variations contribute to most of the human phenotypes, including complex physical traits or disease traits[13,14]. Common polymorphisms include tandemly repeated segments like the minisatellites with 0.1 – 20 kilobase (1,000 nucleotides), and microsatellites with 2 – 100 nucleotides, large copy number variants with 1 kilobase to several megabases, small segmental deletions/insertions/duplications and single nucleotide polymorphisms (SNPs)[14].

SNPs are DNA sequence variations that are caused by the difference of a single nucleotide. This type of variation contributes to about ~ 90% of human genetic variations and occurs in approximately every 500 - 1000 base pairs throughout the human genome[15]. SNPs may occur in coding or non-coding regions of genes. SNPs in coding regions can lead to changes in structure and biological properties of the encoded protein. Meanwhile, SNPs in non-coding regions may affect gene expression levels in an allele-specific manner, for example *MAPK1, FGFR2, LSP1* and *TNRC19* genes [13,15–21].

1.5 Cis-Acting Regulatory Elements

Gene expression is influenced by environmental variation, epigenetic modifications and genetic regulatory elements. One type of genetic regulatory elements is the cis-acting regulatory elements, which are regions of non-coding DNA that regulates the expression of gene on the same chromosome and are often located in close proximity to the regulated gene[22]. Cis-acting regulatory elements are considered to involve regulatory elements such as promoters and enhancers, as well as to be located in intronic regions, upstream of the gene, or hundreds of kilobases away[23].

As previously mention, an example of cis-acting regulatory elements are enhancers that works by binding to activator proteins that trigger DNA bending. Afterwards, these activators interact with co-activators to stimulate chromatin remodelling and histone modification. The activators then bind to mediators, triggering the assembly of RNA polymerase and general transcription factors at the promoter site. Finally, transcription begins (Figure 4).

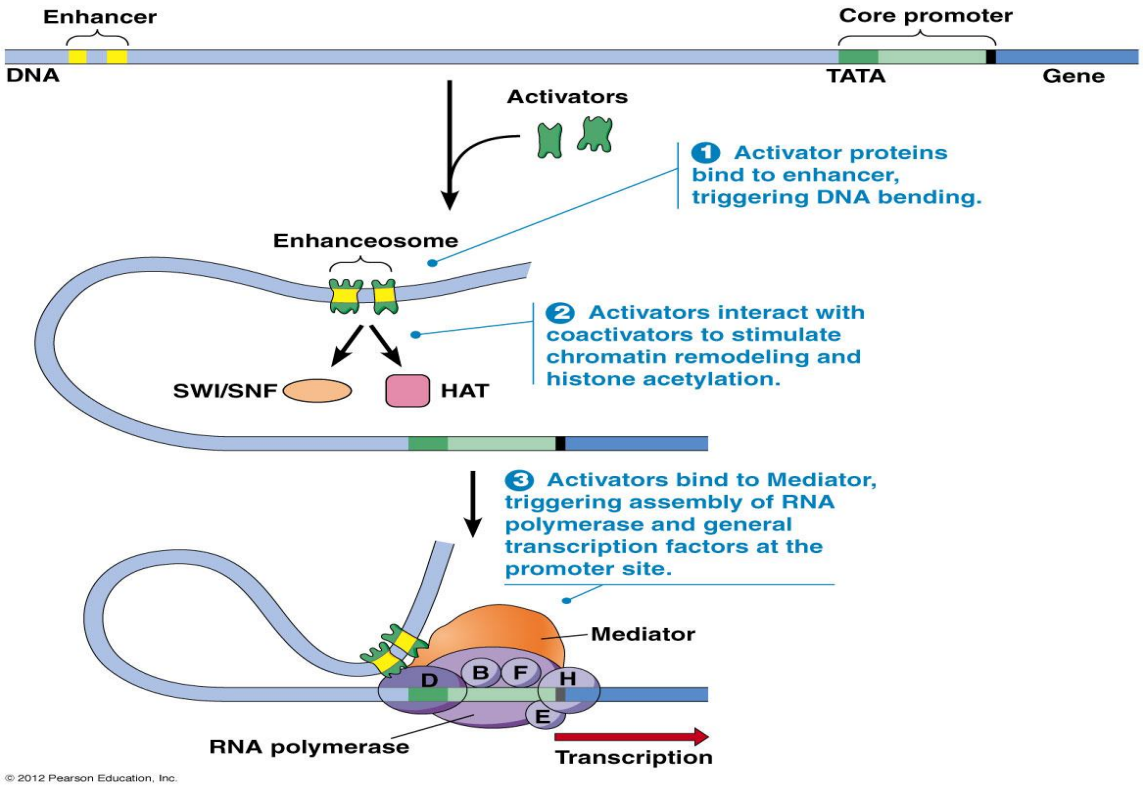


Figure 4 : Mechanism of Cis-Acting Regulated Gene Expression. Pearson Education Inc., 2012.

Cis-acting regulatory variants causes unequal levels of transcription as it favours a specific allele of the gene. This particular allele may contain a specific variant in a proximal promoter that may prevent transcription factor binding, altering the expression of the allelic transcript or vice versa (Figure 5.a). Another possible mechanism could be that a specific variant in a distal enhancer site may prevent combinatorial binding and affect transcriptional levels or vice versa (Figure 5.b).

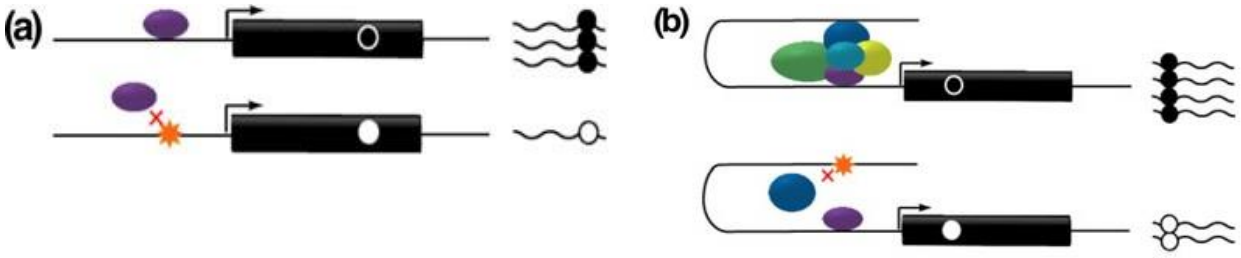


Figure 5 : Differential allelic expression in heterozygous through cis-acting regulatory. SNP in a promoter (a) or enhancer (b) may alter the binding affinity of transcription factors and this affects the level of expression of the alleles shown here as the black and white dots. Modified from Jones et. al., 2011.

Cis-acting regulatory variants can be identified by two different approaches: expression quantitative trait loci (eQTL) and differential allelic expression (DAE). eQTL seeks for association of variations in gene expression between different groups of genotypes. Meanwhile DAE compares the levels of the two transcribed alleles in an individual who is heterozygous for a transcribed SNP (tSNP)[13,22,24].

Several studies done by my research group and others, have shown that most of the variants identified by GWAS for multiple type of cancers are cis-acting regulatory[12,18,20,25,26]. **Therefore, we hypothesise that the remaining risk associated variants identified from GWAS loci outside of genes are likely to be cis-acting regulatory, as well.**

1.6 Differential Allelic Expression

DAE identifies the difference of expression between two alleles of a gene. This is done by comparing the relative expression of the two transcribed alleles in a gene from the same heterozygous sample[24]. In the previous work of my supervisor, Professor Ana Teresa Maia and her colleagues, they have performed a DAE scan of the entire genome in normal breast tissue samples using microarrays and obtained a whole genome map of cis-acting regulatory SNPs in breast tissue (Figure 6). From this study, it was able to identify 7011 (21%) SNPs and 4258 (26%) genes that showed DAE.

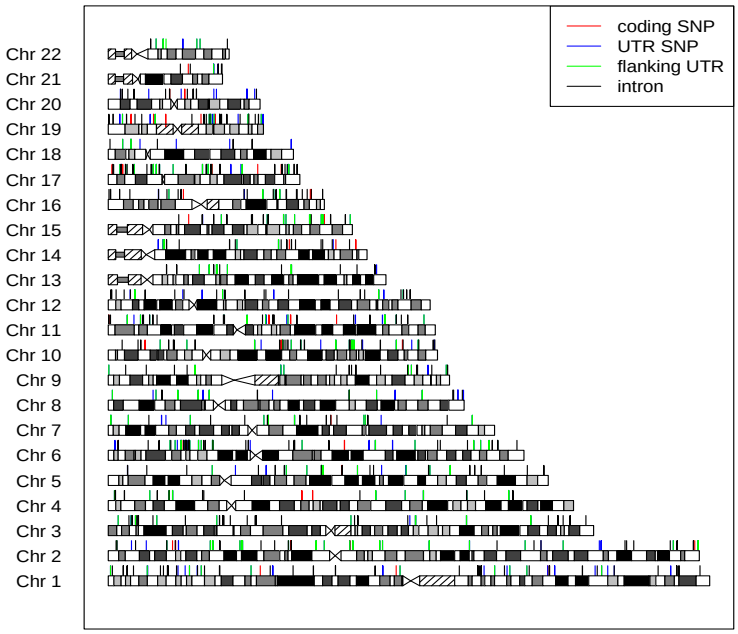


Figure 6 : Genome wide differential allelic expression in breast tissue. Maia. et. al., unpublished.

DAE distribution patterns from the DAE map done by Maia et. al. were consistent with the scenarios presented by Xiao et. al (Figure 7)[22]. These scenarios are attributed to the different levels of linkage disequilibrium (LD) with the base measurement of r^2 and D' between the transcribed SNP (tSNP) and the regulatory SNP (rSNP). LD is the none random correlations among neighbouring alleles, indicating haplotypes inherited from single, ancestral chromosomes[27].

In Scenario 1, the tSNP is in complete LD with the rSNP (r^2=1) and all heterozygotes show DAE. This occurs because there is no recombination between tSNP and rSNP. Hence all heterozygous individuals with tSNP will also be heterozygous for rSNP. Furthermore, only two haplotypes exist in the population and the distribution of DAE can be seen as unidirectional for this scenario (Figure 7.a)[22].

In Scenario 2, the tSNP is in strong but not complete LD with the rSNP (r^2<1, D'=1). In this aspect, heterozygous individuals with tSNP might be heterozygous or homozygous for the rSNP alleles. Therefore three possible haplotypes exist in the population (Figure 7.b)[22].

In Scenario 3, the tSNP and rSNP are in low LD with r^2 and D'<1 hence four haplotypes exist in the population. DAE distribution in this scenario can be seen to be centred around 0 as both of the alleles are equally expressed (Figure 7.c)[22].

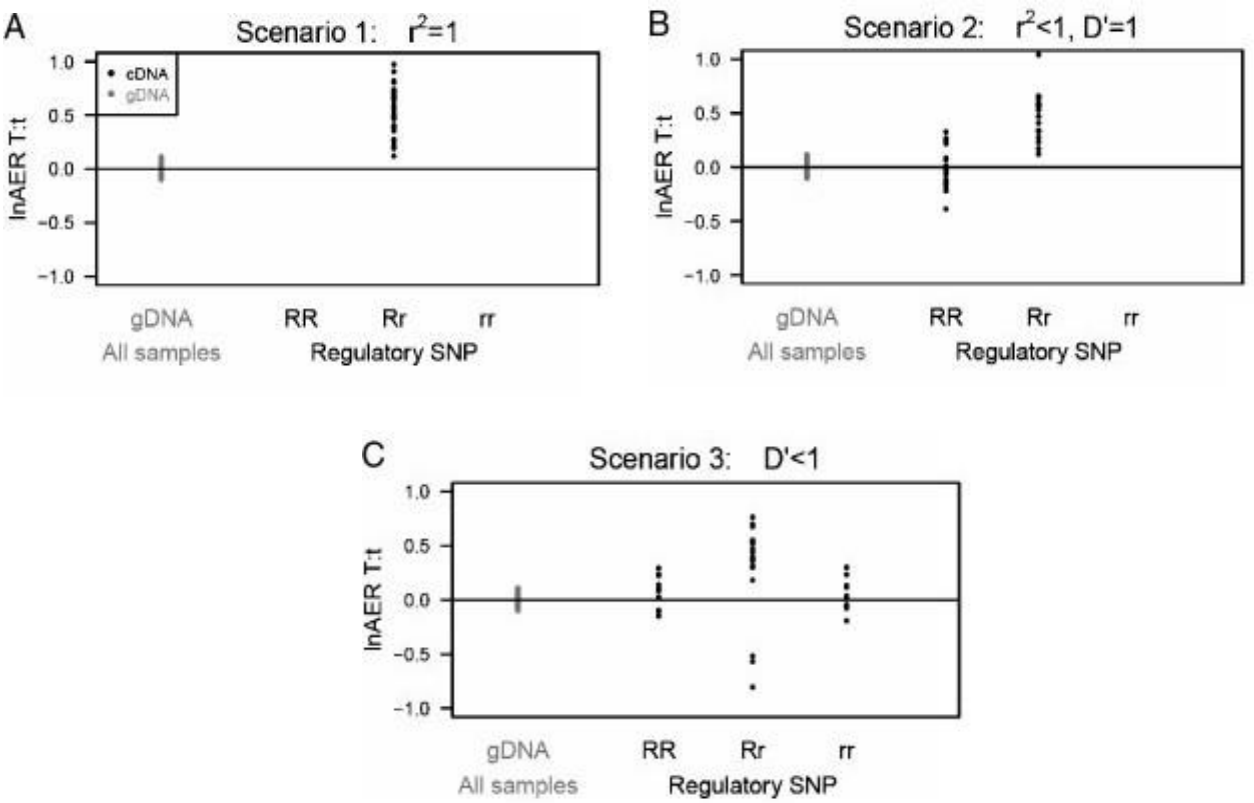


Figure 7: DAE Scenarios for different LD measurements between tSNP and rSNP. Xiao et. al., 2011.

1.7 Preliminary Data

As to date, 94 loci associated with breast cancer susceptibility have been identified and for our group project, we used these GWAS SNPs information and overlapped it with the DAE map by Maia et. al.. (Overlapping data from GWAS SNPs and DAE map was done by another member of the research group, Dr Joana Xavier.) We looked for GWAS and DAE SNPs that were within ± 250 kilobases with each other and had pairwise LD scores of $r^2 > 0.6$. Nineteen loci, with both GWAS associated SNPs and tSNPs displaying DAE, were identified.

Interestingly, the locus 5q14.2, had five tSNPs displaying DAE, three of them in LD with the GWAS SNP rs7707921, which was discovered through meta-analysis performed for 52 GWASes by Michailidou et. al. in 2015[12].

The tSNPs involve are: rs10068160, rs10044824 and rs7733620, located at the gene *ATG10*; rs150934 located at the gene *ATP6AP1L* and rs226202 located at the gene *RPS23*. All tSNPs showed different LD scores with the GWAS SNP (rs7707921) and have uni-directional distribution similar to Xiao et. al. scenario 2, suggesting a strong but incomplete LD, between the regulatory SNP (rSNP) and the transcribed SNP (tSNP) (Figure 8). This distribution also indicated us that the samples that display DAE should be heterozygous for the rSNP and the samples that do not display DAE should be homozygous for the rSNP.

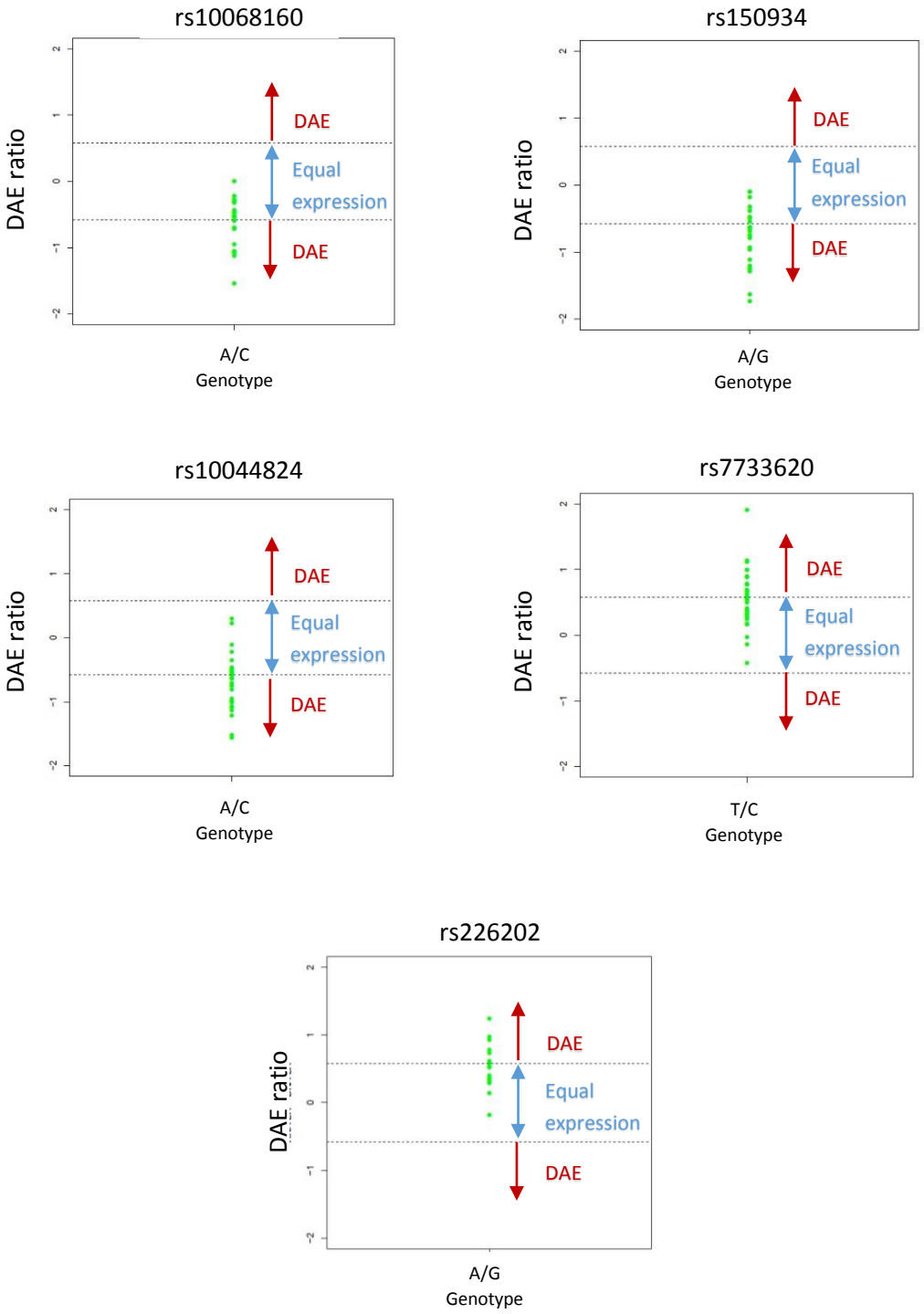


Figure 8 : Five tSNPs with association with GWAS SNP rs7707921. In x-axis it indicate the heterozygous genotype and the y-axis indicates the normalised DAE ratio obtained. Dotted lines delimit the cut-off of preferential allelic expression ratio [Log 2 (1.5) = 0.584].

We hypothesise that the causal variant behind GWAS signal in, rs7707921, is cis-regulatory and the DAE map can help to identify the causal regulatory variant as described in Figure 9. We will use DAE information in order to identify the possible regulatory and risk-causing SNP(s) and then will functionally analyse it. In conclusion, we believe that DAE can be a strong tool to help prioritise GWAS data and to identify causal risk variant(s) and target gene(s) in cancer studies.

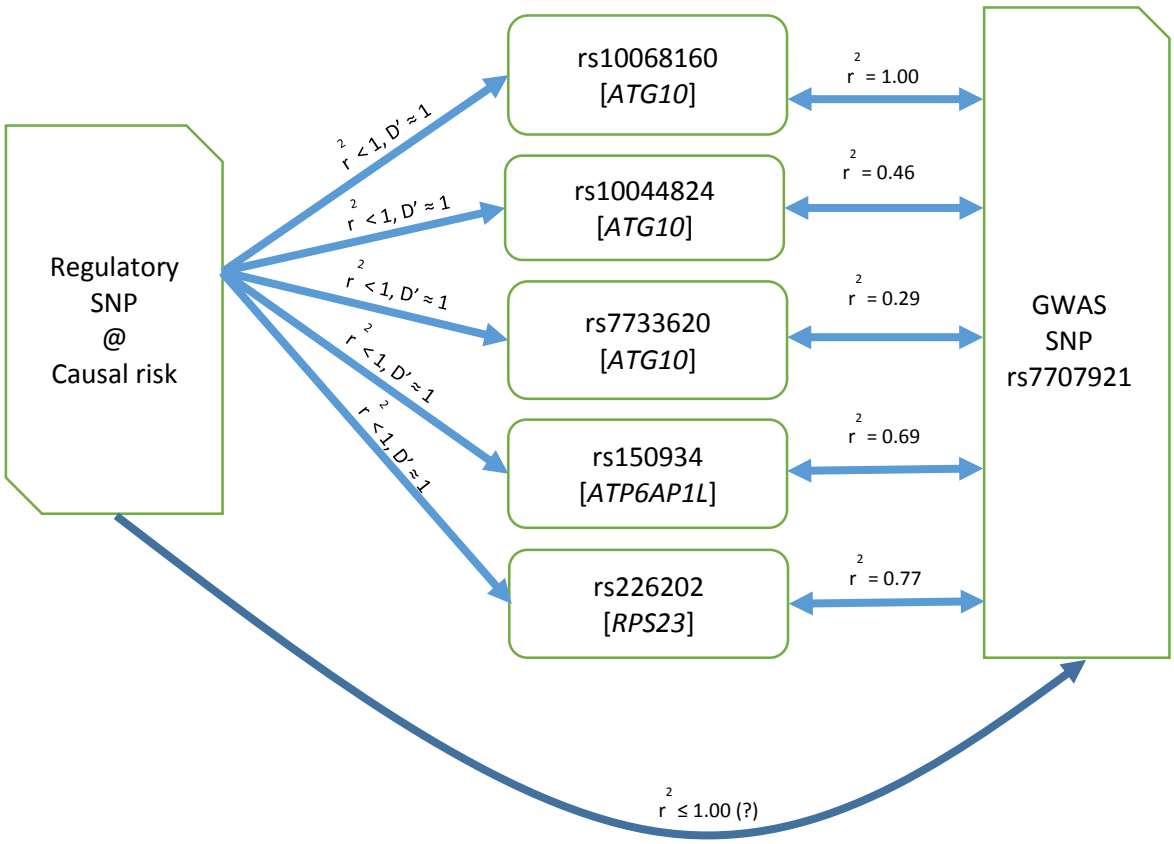


Figure 9 : Association between GWAS SNP, rs7707921 and the correlated five tSNPs from Maia et. al. DAE map.

Chapter 2: Aims

The aims of this study are:

1. To identify candidate cis-acting regulatory SNPs that may be responsible for the association identified at the GWAS meta-analysis SNP: rs7707921. This will be performed by integrating information from breast cancer risk with the DAE in the tSNPs: rs10068160, rs150934, rs10044824, rs226202 and rs7733620.
2. To functionally analyse the regulatory potential of the candidate SNPs identified.

Chapter 3: Material and Method

3.1 Study Samples

In this study, a total number of 290 samples were used. Eighty four samples were from normal breast tissue extracted from women who had undergone reduction mastectomy for reasons not related to cancer. These normal breast tissues were collected at Addenbrooke's Hospital, Cambridge, United Kingdom. An additional 150 and 56 human B cell samples were extracted from anonymous healthy blood donors and cancer patients, respectively. These samples were collected in the Blood Centre at Addenbrooke's Hospital with the approval of Addenbrooke's Hospital Local Research Ethics Committee (REC Reference 04/Q0108/21 and 06/Q0108/221).

DNA extraction were obtained from all samples using conventional SDS/Proteinase K/Phenol method meanwhile total RNA were extracted using TRIzol® method. All extraction procedures were done at University of Cambridge and the extracted RNA used for DAE analysis.

3.2 *In Silico* Annotation of Variants Functional Information

A computational approach was taken in order to identify regulatory SNP associated with breast cancer susceptibility in the 5q14.2 locus. Candidate regulatory SNPs with LD $r^2 \geq 0.4$ with rs7707921, were retrieved from HaploReg V3 database, (http://www.broadinstitute.org/mammals/haploreg/haploreg_v3.php).

HaploReg is a bioinformatics tool used to explore annotations of non-coding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease associated loci. Using linkage disequilibrium information from the 1000 Genomes Project, it allows for the visualisation of linked SNPs and small indels along with their predicted chromatin state, their sequence conservation across mammals and their effect on regulatory motifs[28]. The effect of SNP on regulatory motifs was constructed from Position Weight Matrix (PWM) and scored based on

genomic sequences. PWM is a representation of the DNA binding preferences of transcription factors, used in computational molecular biology and regulatory genomics[29].

Additional functional information was obtained from the following database in order to shortlist candidate regulatory SNPs: the RegulomeDB database, (http://regulomedb.org/), is a database that annotates SNPs with known and predicted regulatory elements in the intergenic regions of human genome. Known and predicted regulatory DNA elements include regions of DNase I hypersensitivity, binding sites of transcription factors and promoter regions that have been biochemically characterised to regulate transcription. The database source include public datasets such as GEO, ENCODE project, Roadmap and published literature[30].

GEO (Gene Expression Omnibus) is a public repository that archives and freely distributes microarray, next-generation sequencing and other forms of high-throughput functional genomics data submitted by the research community[31,32]. ENCODE (Encyclopedia of DNA Elements) is a research project that aims to identify and characterise functional elements in the human genome, which can be accessed through public databases[33]. Roadmap is a database by The NIH Roadmap Epigenomics Mapping Consortium that produce a public resource of epigenomic map for stem cells and primary ex vivo tissues, selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease[34].

RegulomeDB categorises SNPs with functional evidence according to a scoring scheme. SNPs with a score of 1 are likely to affect binding and are linked to expression of a gene target while SNPS with score of 2 are likely to affect binding and score 3 are less likely to affect binding. Lastly, SNPs with scores from 4 to 6 have minimal binding evidence. Scoring of a to f indicates the decrease of functional evidence in a SNP (Annex 1).

The Human Protein Atlas database, (http://www.proteinatlas.org/), provides information on protein expression profiles from normal tissues, cancer tissues, cell lines, subcellular localisation and transcript expression levels[35].

Regulatory landscape for the candidate regulatory SNPs were generated on UCSC Genome Bioinformatics (http://genome-euro.ucsc.edu/index.html) using UCSC Genome Browser on Human Feb. 2009 (GRCH37/hg19) Assembly. This online genome browser hosted by University of California, Santa Cruz (UCSC) is an interactive website that contains several

functional data such as histone modifications, DNAse I hypersensitivity sites, transcription factor ChIP-seq and chromatin state segmentation[36]. Histone modifications regulate chromatin structure and function by processes such as methylation and acetylation. These types of modification are markers for regulatory elements such as enhancers and promoters[37].

In vivo ChIP-seq data were retrieved from Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) database[38], (http://www.broadinstitute.org/ccle/home/54591/), using the database Integrative Genomics Viewer (IGV) version 2.53 analysis tool[39,40]. CCLE provides public access to genomic data, analysis and visualisation for about 1000 cell lines where as IGV is a high performance integrated visualisation of copy number, gene expression, phenotype and other genomic data. Meanwhile ChIP-seq (Chromatin immunoprecipitation sequencing) is a method used to analyse protein interaction with DNA. It combines chromatin immunoprecipitation with DNA sequencing to identify the binding sites of DNA-associated proteins.

3.3 DAE Mapping Analysis

A DAE mapping analysis was performed by plotting the distribution between DAE ratios against the candidate rSNPs genotypes groups (homozygous or heterozygous). A Welch's t-test (unequal variances t-test) was applied to test the equality of DAE means between the homozygous and heterozygous group of samples for each candidate rSNP. The rSNPs genotypes used were retrieved from the previous DAE map study microarrays, as well as imputed (performed by Dr Joana Xavier). Genotype imputation is a process of predicting genotypes that are not directly assayed in individual samples. It is done by having a reference panel of haplotypes at a dense set of SNPs genotyped and a study sample of individuals that have been genotyped at a subset of these SNPs. This approach allows to predict the genotypes at the SNPs that are not directly genotyped in the study sample[41]. DAE mapping analysis was done in order to see if the candidate rSNPs could explain the DAE distribution in each tSNP.

3.4 Genotyping

We did not have genotype information for rs2407153 and rs226198, therefore primers for both candidate SNPs (Annex 2) were designed for the purpose of genotyping study samples using polymerase chain reaction (PCR) method followed by Sanger sequencing. For both rs2407153 and rs226198 rSNPs, 4 ng of genomic DNA in final volume of 20 µL of PCR reaction constituted by KAPA 2G Robust Kit (Kapa Biosystems) and KAPA 2G Fast ReadyMix PCR Kit (Kapa Biosystems) master mix, respectively. PCR conditions were as follows: for KAPA 2G Robust Kit, initial denaturation at 95°C for 3 minutes, denaturation at 95°C for 30 seconds, annealing at 60°C for 30 seconds, extension at 72°C for 30 seconds and final extension at 72°C for 1 minutes before cooling down to 4°C. Denaturation, annealing and extension steps were repeated for 30 cycles. Meanwhile for KAPA 2G Fast ReadyMix PCR Kit, initial denaturation at 95°C for 3 minutes, denaturation at 95°C for 15 seconds, annealing at 60°C for 15 seconds, extension at 72°C for 1 seconds and final extension at 72°C for 1 minutes before cooling down to 4°C. Denaturation, annealing and extension steps were repeated for 30 cycles. PCR was performed using C1000 Touch Thermal Cycler (Biorad).

PCR amplification was confirmed by performing size separation electrophoresis in a 1.5% agarose gel at 100 V for 60 minutes. Exo/SAP Go – PCR Purification Kit (Grisp) was used for PCR products clean up before the concentration of the PCR products was measured using the Nanodrop 2000c Spectrophotometer (Thermo Scientific). Afterwards, the PCR products were diluted to 80 ng/µl and 5 µl were sent for sequencing with Sanger Method.

3.5 Cell Lines

HCC1954 and T-47D cell lines were maintained in RPMI medium supplemented with 10% of foetal bovine serum and 1% of penicillin. Meanwhile MDA-MB-231 cell line was maintained in DMEM medium supplemented with 10% of foetal bovine serum and 1% of penicillin. All HCC1954, T-47D and MDA-MB-231 cell lines were obtained from our collection. HCC1954, T-47D and MDA-MD-231 are all cancer cell lines with HCC1954 and MDA-MD-231 being estrogen receptor negative (ER-) meanwhile T-47D is an estrogen receptor positive (ER+).

Other cell lines analysed during this master project, either *in silico* or experimentally are: human mammary epithelial cells (HMEC), human mammary fibroblasts (HMF), oestrogen receptor negative cancer mammary epithelial cells (MCF10A) and oestrogen receptor positive cancer mammary epithelial cells (MCF-7).

3.6 Nuclear Protein Extraction

Nuclear protein extraction was prepared using the NE-PER Nuclear and Cytoplasmic Extraction Reagents kit (Thermo Scientific, Life Technologies) following the protocol stated by the manufacturer. Nuclear protein extract concentration was measured using Nanodrop 2000c Spectrophotometer (Thermo Scientific) in accordance to the manufacturer's instructions.

3.7 Electrophoretic Mobility Shift Analysis

Electrophoretic Mobility Shift Analysis is an approach that allows the detection of protein and nucleic acid interactions[42]. This method is based on electrophoretic mobility shift between protein and nucleic acid binding complexes using non-denaturing polyacrylamide gel[42,43].

A double strand oligonucleotide containing a putative binding sequence is labelled with a non-radioactive marker, biotin. This labelled oligonucleotide is then mixed with nuclear extract in order for them to form protein-nucleic acid binding complexes. If protein binds to the labelled oligonucleotide, the binding complex will migrate slower than the unbound oligonucleotide through the polyacrylamide gel, creating a shift.

Competitive binding assay is performed by adding unlabelled oligonucleotides in order to test the binding specificity of a protein to the target sequence[43]. Subsequently, antibody binding or super-shift assay can be performed in order to identify the protein present in the protein-nucleic acid complex[44]. This binding complex consist of antibody, protein and nucleic acid will

cause an even lower mobility of protein – nucleic acid complex resulting in a super shift on the gel (Figure 10).

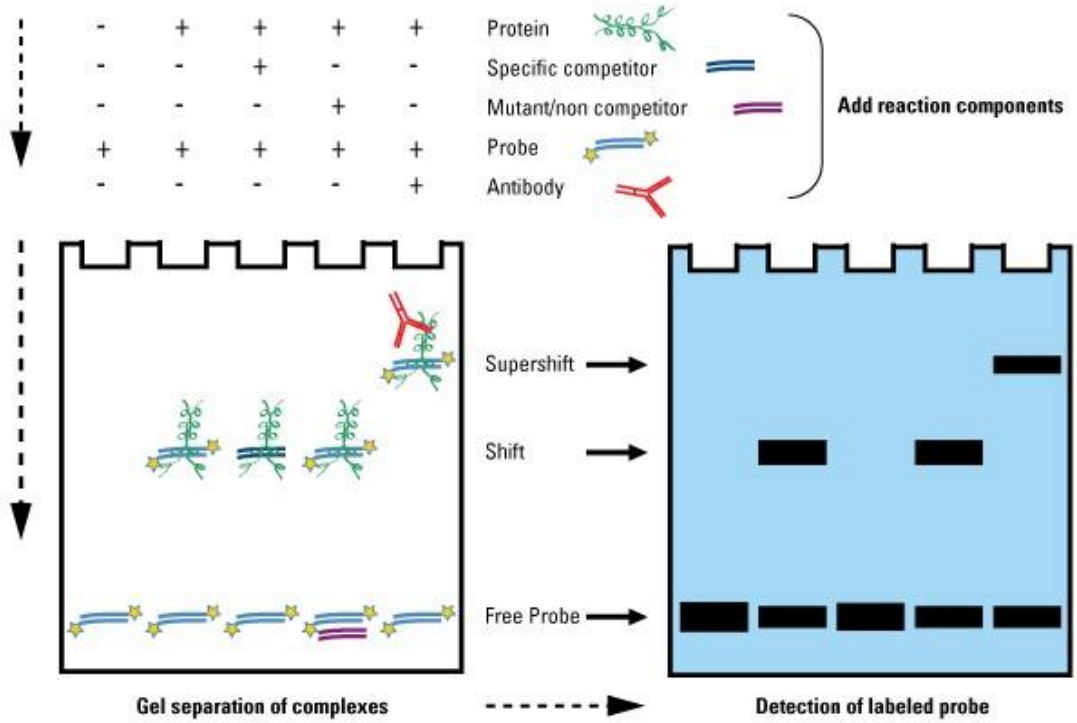


Figure 10: Overview of EMSA method. On the left is the reaction assay scheme for EMSA. On the right is the expected EMSA band scheme seen on gel. Life Technologies Inc., 2015.

3.7.1 Oligonucleotide Labelling

Oligonucleotides with 35 base pairs of DNA sequence neighbouring the rs2406909 SNP was designed (Annex 2) and a previously known oligonucleotide with a sequence complementary to *FGFR2* locus, that binds to Oct-1 and CEBPβ proteins, was used as a positive control[45]. Oligonucleotides labelling was done by separately incorporating biotin to the 3' end of each strand of complementary oligonucleotide. Afterward these complementary oligonucleotides were annealed to each other. Lastly, labelling efficiency was checked after the completion of oligonucleotides annealing process. Oligonucleotides labelling was performed using Biotin 3' End DNA Labelling Kit (Life Technologies) in accordance to manufacturer's instructions. For labelling

efficiency, dot blot using hand spotting protocol was chosen and the detection was done using the Chemiluminescent Nucleic Acid Detection Module (Life Technologies).

3.7.2 Protein-Nucleic Acid Binding and Competition Assay

All EMSAs were performed using Light Shift Chemiluminescent EMSA Kit (Life Technologies). The EMSA for protein-nucleic acid binding complexes detection were performed separately for the two alleles of each SNP. EMSA was carried out with 20 fmol biotin 3' end labelled oligonucleotides, which were incubated for 15 minutes at room temperature with 15 µg nuclear extract in a final volume of 20 µL binding buffer (20 mM Hepes, 0.1 mM $ZnCl_2$, 10 % glycerol, 1 mM DTT, 1x protein inhibitor and 10 ng/µL poly (dI.dC)). A 4-20% TBE Gel (Life Technologies) polyacrylamide gradient gel was used to migrate the samples at 100 V for 2 hour. The protein-nucleic acid complexes were then transferred from the gel onto a nylon membrane using a Mini Trans-Blot Electrophoretic Transfer Cell (Biorad) at 80 V for 1 hour. Crosslink and detection were performed in accordance to manufacturer's instructions.

Following the results obtained in the protein-nucleic acid binding assay, the SNP alleles that showed clear indication of protein binding were further analysed with EMSA competition assay. The competition assays were performed by adding different concentrations of unlabelled oligonucleotide for the alleles of interest. The remaining EMSA protocol was the same as described above for protein-nucleic acid binding assay.

3.8 Haplotypes Analysis

Haplotypes analysis was performed using genotypes information for all tSNPs and rSNPs except for rs226198. SNP rs226198 did not have imputation genotypes as well as proxy SNP in complete LD with it. Therefore, we performed sample genotyping for rs226198. We also did not have genotypes information for GWAS SNP. However, a proxy SNP rs6888977, that is in complete LD ($r^2 = 1$) with GWAS SNP was identified. Haploview software was used to analyse haplotype structure for tSNPs, rSNPs and proxy GWAS SNP[46].

3.9 Gene Expression

Real-time PCR allows PCR amplification and detection into a single step, hence eliminating the need to detect products using gel electrophoresis and enables the PCR products to be quantified. Real-time PCR uses fluorescent dyes to detect the accumulation of PCR products during the exponential phase of the reaction.

Amplification plot in a real-time PCR is performed through the detection and quantification of a fluorescent reporter signal. Baseline for the real-time PCR is established first, before the threshold line and cycle thresholds (Ct) are calculated, in the beginning of the exponential phase. Threshold line is the level of detection, where a reaction reaches a fluorescent intensity, above the established baseline. Ct is the point of intersection between the amplification curve and the threshold line[47], and is a relative measure of the concentration of the target product, in the PCR reaction. In the exponential phase it is assumed that the signal intensity is in direct proportion to the amount of PCR product and, therefore, the amount of fluorescence is registered at each cycle and measured.

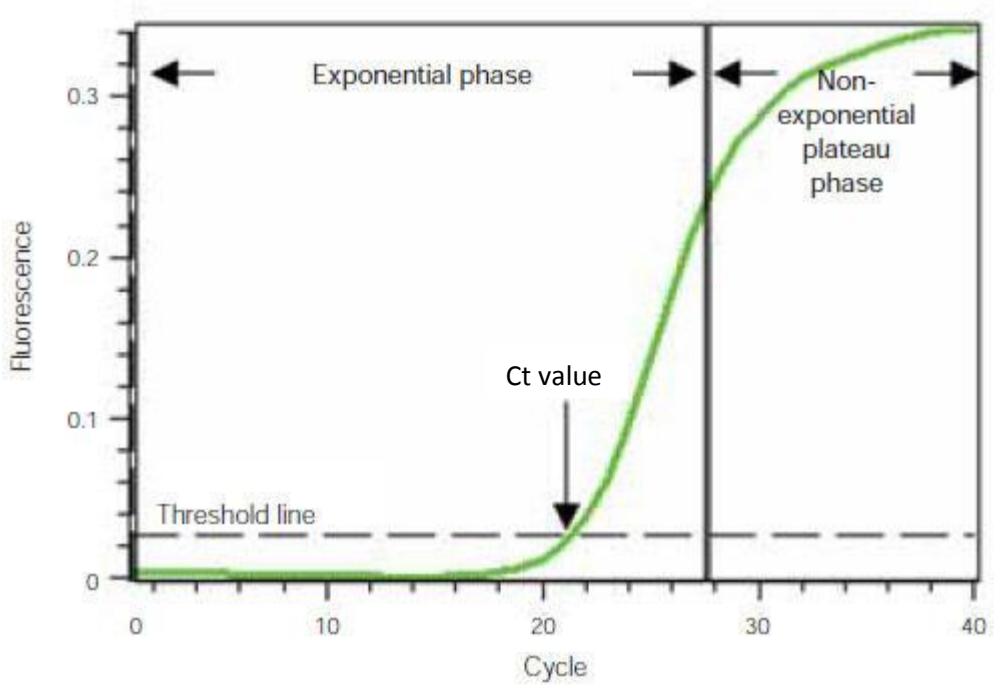


Figure 11: Real time PCR detection of product in exponential phase. Ct values are calculated by determining the point at which the fluorescence exceeds a threshold limit.

There are several types of detection chemistry used in qPCR, namely: DNA binding dyes, hybridization probes, hairpin probes and Taqman probes. In this study, we used the Taqman probes that utilizes the 5' nuclease activity of DNA polymerase to hydrolyse a hybridization probe bound to its target product. The probe emits a fluorescent signal when cleaved based on the fluorescent resonance energy transfer (FRET) principle. The probe 3' end is non-extendable and is dual-labelled, with a reporter fluorochrome such as FAM or HEX and a quencher fluorochrome. The probe is designed to anneal to the target sequence internally of the primer during the annealing and extension phase of the PCR reaction. When the probe is in its intact and free form, it does not emit any fluorescence because the fluorescence of the reporter dye is absorbed by the quencher dye by the principle of FRET. However, upon extension of the nascent strand, the probe become degraded by the activity of 5' to 3' exonuclease activity of the Taq polymerase, resulting in an increase of reporter fluorescent emission which is proportional to the increase in PCR product (Figure 12)[48].

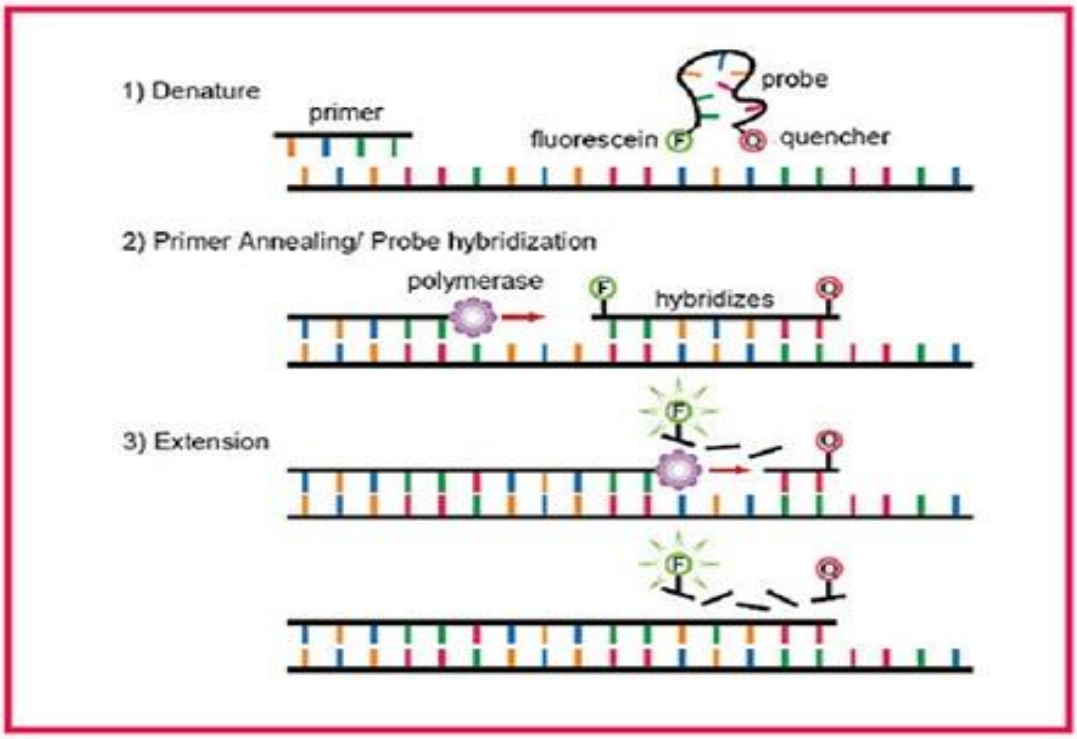


Figure 12: Detection chemistry in qPCR by hydrolysis of Taqman probe. In free and intact form, the Taqman probe does not emit fluorescence. In the PCR reaction, the probe anneal to target sequence and then get hydrolysed by 5' to 3' exonuclease activity of the Taqman polymerase which later cleavage the reporter dye away from the quencher dye. This enable the reporter dye to emit fluorescent by the principle of fluorescent resonance energy transfer (FRET).

RPS23 gene expression levels were assayed using quantitative real-time polymerase chain reaction (qPCR) method (Hs01374150_g1 probe) with 29 complementary DNA (cDNA) samples prepared from normal breast tissue, already existing in the laboratory. The *GAPDH* gene was used as housekeeping gene and its expression was also measured for the 29 samples. A serial dilution using a DNA sample from a MCF-7 cell line was included for both *RPS23* and *GAPDH* assays in order to construct a standard curve for gene expression quantification. Serial dilutions were set at 1:1, 1:2, 1:10, 1:20, 1:100, 1:1,000 and 1:10,000. Five ng/µL of cDNA of each sample were reconstituted to final volume of 5 µL by a master mix consisted of Taqman Gene Expression Assays ((20X, Applied Biosystems), Kapa Probe Fast Universal qPCR Kit ((2X), Kapa Biosystems) and DNase/RNase free water (Life Technologies). Each sample (for *RPS23*, *GAPDH* assays, as well as serial dilutions) was assayed in triplicate and samples were assayed on the same 384 well plate.

Ct values were obtained from Bio-rad CFX Manager Software and analysed on Microsoft Excel 2013. Calculation of *RPS23* gene expression levels for each sample were obtained from averaged Ct values and normalized according to GADPH measurements, for each sample individually. Total gene expression data for *ATG10 and*, *ATP6AP1L* have been previously measured with Illumina HumanHT-12 expression beadchips for 27 samples in common with the DAE map (Probes ID: ILMN_2206098 and ILMN_1755990 for *ATG10* and *ATP6AP1L*, respectively).

For eQTL analysis, *RPS23, ATG10* and *ATP6AP1L* expression levels were plotted against genotypes at each candidate rSNP. For all rSNPs, a logistic regression using log-additive model was applied to assess the association of genotypes with total gene expression levels. Statistical analysis and graphics were performed using R software.

Chapter 4: Results

4.1 Identification of Candidate rSNPs in the 5q14.2 Locus

One of the objectives of this study was to identify cis-acting regulatory SNPs that may be responsible for the breast cancer risk association identified for SNP rs7707921[12]. This was performed by analysing the linkage disequilibrium between the GWAS SNP with the five tSNPs in the *ATG10-RPS23-ATP6AP1L* locus: rs10068160, rs150934, rs10044824, rs226202 and rs7733620. Candidate rSNPs list was generated by selecting SNPs with LD, $r^2 \geq 0.4$ with the GWAS SNP, based on DAE distribution showing Scenario 2, according to Xiao et. al., for all tSNPs[22]. This scenario, suggests a strong, but not complete, LD between the regulatory SNP and the transcribed SNP.

We retrieved 125 candidate rSNPs with the established LD cut-off from HaploReg V3 database. Additional functional data were sought from RegulomeDB and The Human Protein Atlas databases. From the 125 candidate SNPs, six SNPs were found to have evidence of regulatory functions in breast tissue or breast cell lines, as shown in Table 1. Information obtained from RegulomeDB showed that each of these six candidate rSNPs have more than one functional evidence such as being an eQTL, having transcription factor (TF) binding, matched TF motif or DNAseI hypersensitivity.

Table 1: RegulomeDB scores for the candidate regulatory SNPs.

SNP	RegulomeDB Score
rs226198 rs17247678	1a
rs2406909 rs10036937	1b
rs2407153	1d
rs6880209	1f

RegulomeDB categorises SNPs with functional evidence according to a scoring scheme. SNPs with a score of 1 are likely to affect binding and are linked to expression of a gene target while SNPS with score of 2 are likely to affect binding and score 3 are less likely to affect binding. Lastly, SNPs with scores from 4 to 6 have minimal binding evidence. Scoring of a to f indicates the decrease of functional evidence in a SNP.

4.2 Analysis of Candidate rSNPs in the 5q14.2 Locus

The six candidate rSNPs were categorised into 4 regions in the regulatory landscape in the 5q14.2 locus, as visualised in Figure 13, using the UCSC Genome Browser. Region 1 contained one candidate rSNP rs24060909. In region 2, two candidate rSNPs rs10036937 as well as rs2407153. In region 3, also two candidate rSNPs rs226198 and rs6880209. Finally region 4 had candidate rSNP rs17247678 (Figure 13).

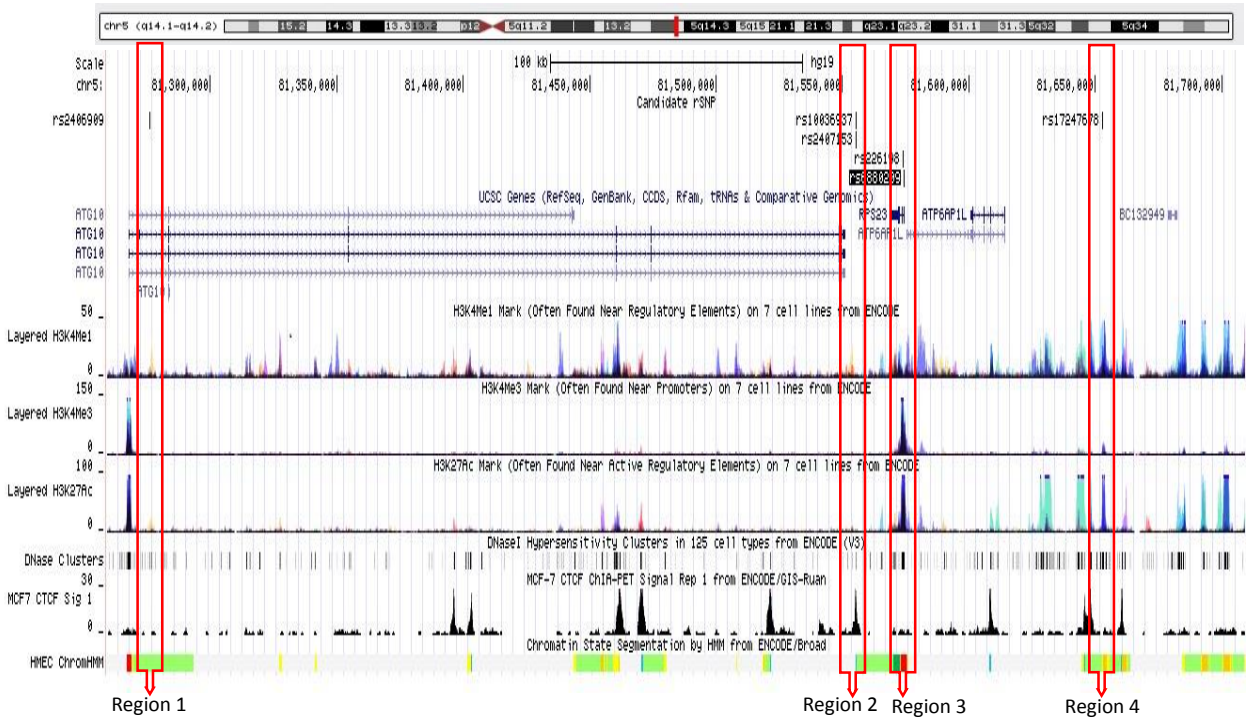


Figure 13: Genomic view of the 5q.14.2 locus with functional regulatory evidence. From the top to bottom of the figure are shown the candidate rSNPs, the RefSeq genes mapped to the area of interest around the 5q14.2 locus, information on histone modifications, DNase I hypersensitivity clusters, CTCF binding region in MCF-7 and chromatin modifications in HMEC according to the Genome Browser (http://genome-euro.ucsc.edu/index.html).

Genomic view of the 5q14.2 locus showed information on histone modifications such as H3K4Me3, H3K4Me1 and H3K27AC, which are markers for promoters, enhancers and active regulatory elements, respectively. There was also information on DNase I hypersensitivity site (DHS), which is a region that organises DNA structure and acts as regulator of transcription by enabling or restricting protein binding[49,50]. CTCF in MCF-7, a region known to mediate chromatin loop and causes accessibility to regulatory elements region either to be activated or blocked, hence influencing transcriptional regulation[51]. Lastly, chromatin modifications in HMEC.

4.2.1 Region 1: Candidate rSNP rs2406909

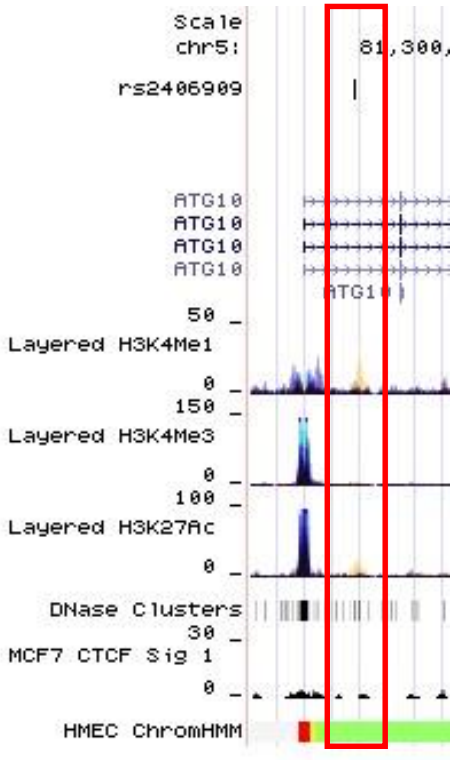


Figure 14: Genomic view of the region 1: rSNP rs2406909 with functional evidence. From top to bottom of the figure is shown the candidate rSNP, the RefSeq gene mapped to the area of interest around the 5q14.2 locus, information on histone modifications, DNase I hypersensitivity site, CTCF binding region in MCF-7 and chromatin modifications in HMEC according to the Genome Browser (http://genome-euro.ucsc.edu/index.html). This candidate rSNP overlaps DHS region.

In region 1, candidate rSNP rs2406909 was found to overlap a DHS. Results from our DAE mapping analysis showed that this candidate rSNP rs2406909 was associated with DAE at the tSNPs rs10044824 (p-value = 3.90 x- 10^{-4}) and rs7733620 (p-value = 3.97 x 10^{-6}) (Figure 15), both located in *ATG10*. From the DAE mapping analysis, we observed that when individuals are homozygous for this candidate rSNP they show equal levels of transcription of the two alleles of the tSNP. Meanwhile individuals heterozygous for this SNP display DAE, suggesting this rSNP as a potential rSNP.

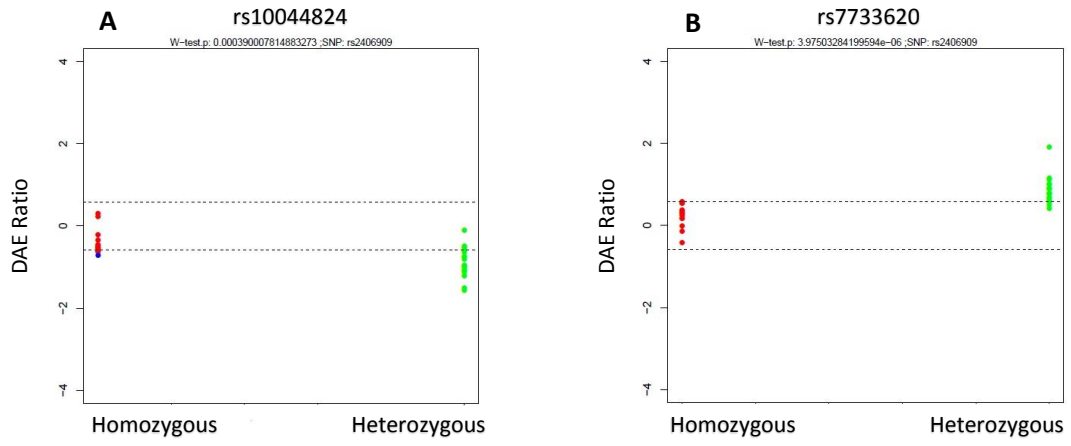


Figure 15: DAE mapping analysis for rSNP rs2406909. The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples observed in between the dotted lines have equal expression hence not causing the effect of DAE meanwhile samples observed above or below the dotted lines have unequal expression hence causing the effect of DAE. DAE pattern at the tSNP are identified by grouping the rSNP sample genotypes. [DAE ratio, log 2 (1.5) = 0.584]. (A) rSNP rs24006909 with tSNP rs10044824. (B) rSNP rs2406909 with tSNP rs7733620.

As rs2406909 was located in a regulatory element region, we searched for evidence of TF binding. Position Weight Matrix (PWM) data analysis (from the Haploreg database) showed that several predicted TFs (HDX, Nanog, Pou2f2, TEF, YY1 and p300) with different mass weight, ranging from 300 kDa to 30 kDa, have different binding affinity to the two alleles of this SNP, as shown in Table 2. We were not able to obtain any ChIP-seq information from the CCLE database for this candidate rSNP. Therefore, we performed EMSA analysis in order to verify if there was any *in vitro* TF binding potential with our rSNP rs2406909, as suggested in PMW analysis.

Table 2: Predicted transcription factor binding and protein mass weight, for candidate rSNP, rs2406909, in Haploreg V3 database.

Position Matrix Weight (PWM)	Reference Allele	Alternative Allele	Protein Mass Weight (kDa)
HDX	9.4	12.8	77
Nanog	11.8	5.3	35
Pou2f2	12.2	6.5	51
TEF	1.5	13.5	33
YY1	10.9	2.3	45
p300	6.6	11.6	264

Before performing EMSA analysis, biotin labelled oligonucleotide efficiency estimation for the candidate rSNP's alleles was determined to be at 100% (Annex 3). In Figure 16, are shown the results for an EMSA for rs2406909 using protein nuclear extract from the breast cancer cell line MCF-7. Lane 1 corresponds to oligonucleotides of *FGFR2* that was used as a positive control, meanwhile lanes 2 and 3 contain oligonucleotides that have the specific T and C allele sequences, respectively. From the gel image, both showed a shift, suggesting that both bind to protein existing in the MCF-7 nuclear extract. Based on the reference of protein marker, the size of the shift corresponded to mass weight ranging from 63 to 75 kDa.

Competition assays were performed, showing in lanes 4 until 14. Lanes 4 to 9 correspond to labelled oligonucleotide containing the T allele, meanwhile lanes 9 to 14 correspond to labelled oligonucleotide containing the C allele. Competition assays were done by adding different concentrations of unlabelled oligonucleotides. These were set at the same concentration as the labelled oligonucleotide, 33-fold higher and 100-fold higher, except for lanes 10 and 11, where the competition concentrations were set at 33-fold higher and 100-fold higher. Lanes 4, 5, 6, 10 and 11 correspond to T allele competition meanwhile lanes 7, 8, 9, 12, 13 and 14 correspond to C allele competition.

In lanes 4, 7 and 12, corresponding to the same concentration of labelled and unlabelled oligonucleotides, we can see that the bands are weaker compared to the lanes that do not have competition (lane 2 and 3). This suggests that protein-nucleic acid binding truly occurred for both alleles, as both labelled and unlabelled oligonucleotides bind to protein. This suggestion is further supported by the absence of bands in lanes that contain 33-fold higher and 100-fold higher unlabelled oligonucleotides.

The EMSA was repeated with the same conditions using a different breast cancer cell line, MDA-MD-231, which does not express oestrogen receptor (ER-). The EMSA result is shown in Figure 17, where we can see bands in lanes 2 and 3 that correspond to both the T and C alleles labelled oligonucleotides, respectively. We can also see bands in lanes 3, 7 and 12 even though it contained competitor oligonucleotides. However, as the concentration of competitor increased, the bands gradually disappeared. This suggests that the protein binding has the same affinity for both alleles, even in different cell lines (ER+ vs ER-). Since there was protein binding

for both alleles, this does not explain the DAE. Therefore, this candidate could not be the SNP responsible for the DAE observed at tSNPs rs10044824 (*ATG10*) and rs7733620 (*ATG10*). Therefore, was excluded from further analysis.

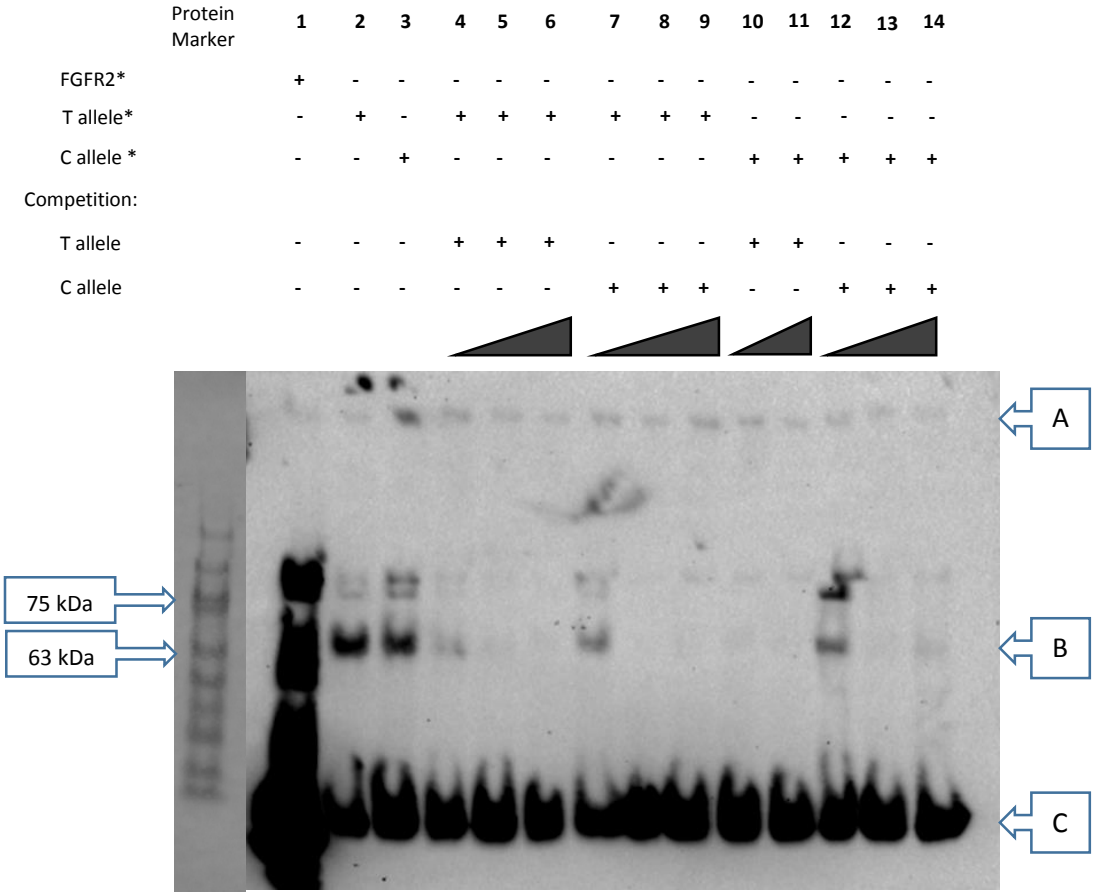


Figure 16: In vitro protein-nucleic acid binding and competition binding studies. EMSA analysis was performed using MCF-7 nuclear extract. Lane 1 corresponds to oligonucleotides of FGFR2 that was used as a positive control, meanwhile lanes 2 and 3 correspond to oligonucleotides that have the specific T and C allele sequences, respectively. Lanes 4 to 9 correspond to labelled oligonucleotide containing the T allele, meanwhile lanes 9 to 14 correspond to labelled oligonucleotide containing the C allele. For competition assays, unlabelled oligonucleotide in lanes 4 to 14 were set at the same concentration as the labelled oligonucleotide, 33-fold higher and 100-fold higher, except for lanes 10 and 11, where the competition concentrations were set at 33-fold higher and 100-fold higher, respectively. Lanes 4, 5, 6, 10 and 11 correspond to T allele competition meanwhile lanes 7, 8, 9, 12, 13 and 14 correspond to C allele competition. A indicate the top of the band (well). B indicate the specific bands for protein-nucleic acid binding. C indicate the free unbound oligonucleotides.

** represent the labelled oligonucleotide.*
+ represent the present of oligonucleotide in the reaction assay.
- represent the absent of oligonucleotide in the reaction assay.
◢ represent the increase of oligonucleotide concentration.

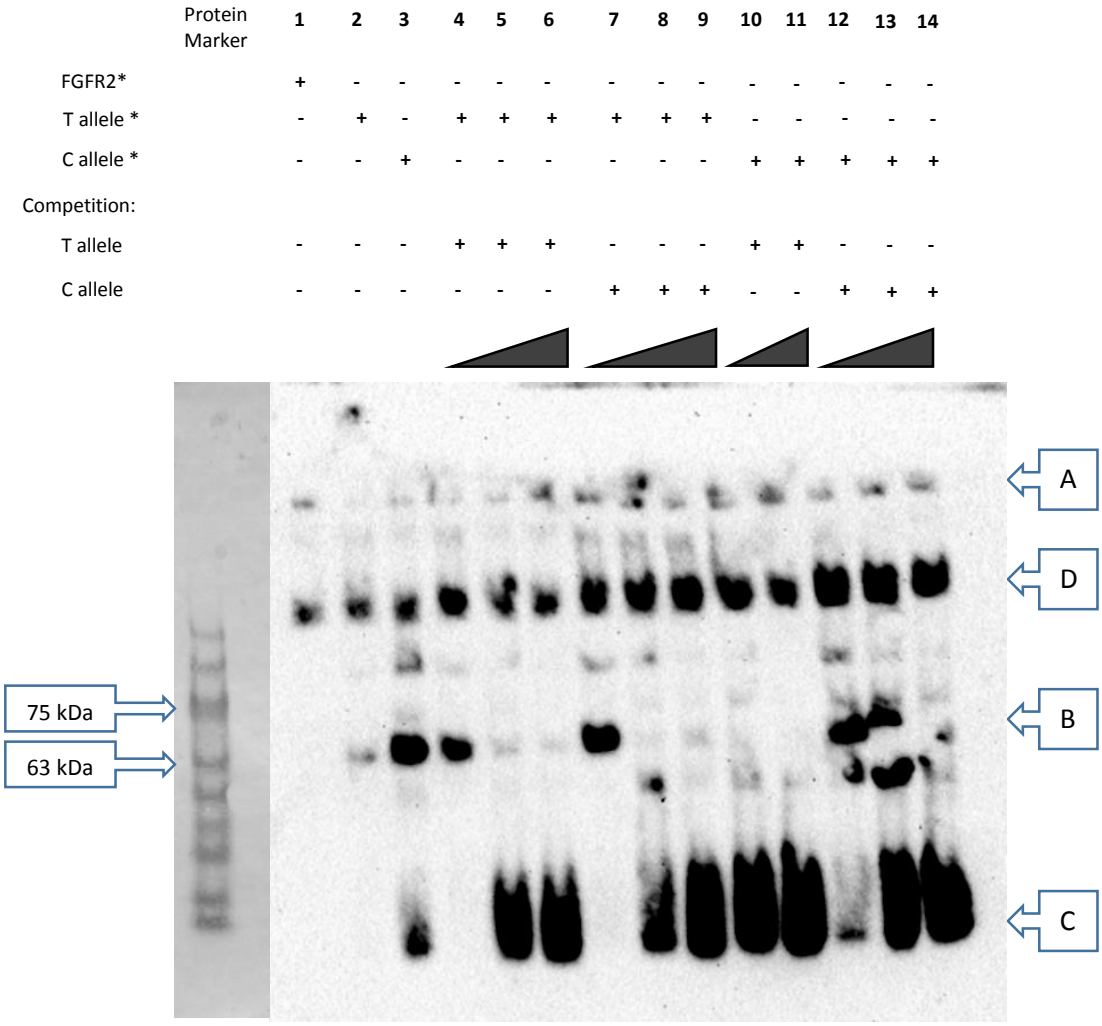


Figure 17: In vitro protein-nucleic acid binding and competition binding studies. EMSA analysis was performed using MDA-MB-231 nuclear extract. Lane 1 corresponds to oligonucleotides of FGFR2 that was used as a positive control, meanwhile lanes 2 and 3 correspond to oligonucleotides that have the specific T and C allele sequences, respectively. Lanes 4 to 9 correspond to labelled oligonucleotide containing the T allele, meanwhile lanes 9 to 14 correspond to labelled oligonucleotide containing the C allele. For competition assays, unlabelled oligonucleotide in lanes 4 to 14 were set at the same concentration as the labelled oligonucleotide, 33-fold higher and 100-fold higher, except for lanes 10 and 11, where the competition concentrations were set at 33-fold higher and 100-fold higher, respectively. Lanes 4, 5, 6, 10 and 11 correspond to T allele competition meanwhile lanes 7, 8, 9, 12, 13 and 14 correspond to C allele competition. A indicate the top of the band (well). B indicate the specific bands for protein-nucleic acid binding. C indicate the free unbound oligonucleotides.

** represent the labelled oligonucleotide.*
+ represent the present of oligonucleotide in the reaction assay.
- represent the absent of oligonucleotide in the reaction assay.
◢ represent the increase of oligonucleotide concentration.

4.2.2 Region 2: Candidate rSNPs rs10036937 and rs2407153

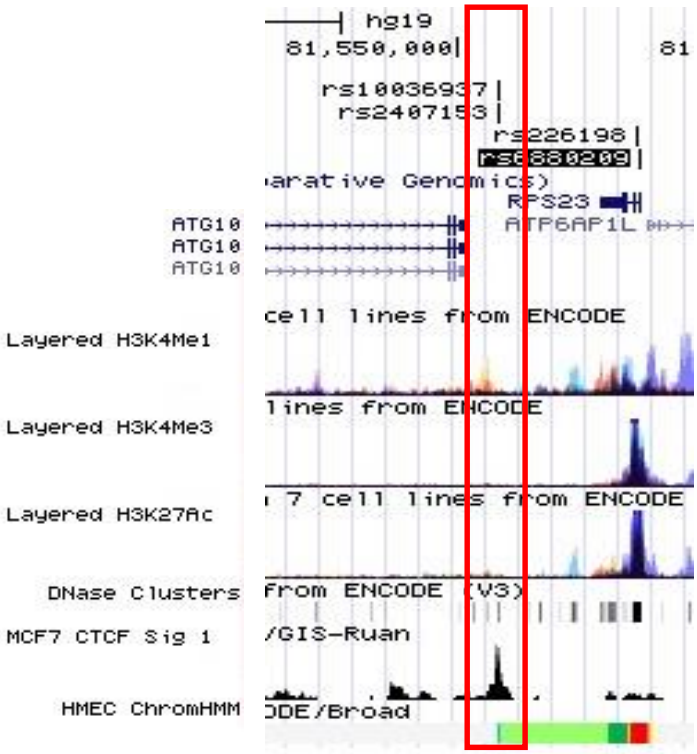


Figure 18: Genomic view of the region 2: rSNP rs10036937 and rs2407153 with functional evidence. From top to bottom of the figure are shown the candidate rSNPs, the RefSeq genes mapped to the area of interest around the 5q14.2 locus, information on histone modifications, DNase I hypersensitivity clusters, CTCF binding region in MCF-7 and chromatin modifications in HMEC according to the Genome Browser (http://genome-euro.ucsc.edu/index.html). These candidate rSNPs overlap DHS and CTCF region.

In region 2, both candidate rSNPs rs10036937 and rs2407153 were found to overlap a CTCF binding region as well as DHS region. This is an indication that this region is under the influence of chromatin looping, as well as activation, or restriction, of protein binding by DHS.

For candidate rSNP rs10036937, we observed in DAE mapping analysis that there was no significant association with tSNP rs7733620 (p-value = 0.081) located at *ATG10* and tSNP rs226202 (p = 0.542) located at *RPS23*. As for rSNP candidate rs2407153, after genotyping 25 samples, we verified that there was a significant association with tSNP rs10044824 (*ATG10*) (p-value = 1.23×10^{-5}) (Figure 19). However, no significant association for tSNP rs7733620 (*ATG10*) with p-value = 0.379 or tSNP rs226202 (*RPS23*) with p-value = 0.517 (Annex 4).

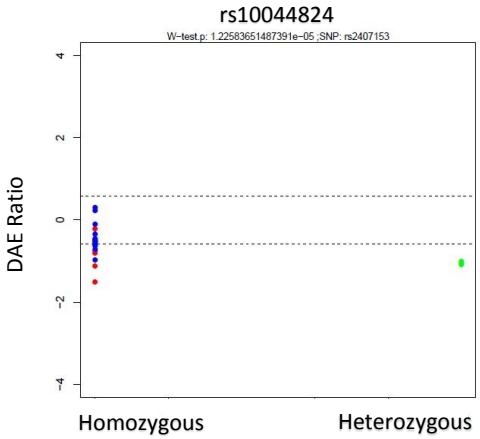


Figure 19: DAE mapping analysis for rs2407153. The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples observed in between the dotted lines have equal expression hence not causing the effect of DAE meanwhile samples observed above or below the dotted lines have unequal expression hence causing the effect of DAE. DAE pattern at the tSNP are identified by grouping the rSNP sample genotypes. [DAE ratio, log 2 (1.5) = 0.584]. This is an *rSNP rs2407153* with *tSNP rs10044824.*

We proceeded with seeking for evidence of TF binding and from the CCLE database, we were able to obtain evidence of CTCF binding to candidate rSNP rs10036937 in MCF-7, HMF and HMEC cell lines. In the MCF-7 cell line, there was CTCF binding data from several ChIP-seq experiments and most of the reads count were higher than 10, with the maximum coverage at 24. The CTCF binding in MCF-7 only occurred to the C allele. As for HMF and HMEC cell lines, we saw preferential CTCF binding to the C allele compared to the G allele. However, for HMF, we only identified data from two ChIP-seq experiments, with preferential CTCF binding to the C allele, also with reads count of 5 and 14 (maximum coverage at 23). Meanwhile in HMEC, total reads count for all experiments were less than 10 with preferential binding reads count between 2 and 4 (Annex 5). Published data also revealed that rs2407153 is bound by CTCF in the MCF-7 cell line. However, all the ChIP-seq experiments reads count were below 3 (Annex 5).

Candidate rSNP rs2407153 showed a significance for association at tSNP rs10044824 (*ATG10*), in DAE mapping analysis consisted of 25 genotyped samples. However, both rs2407153 and rs10036937, lacked evidence for regulatory elements and had low reads count for CTCF binding. Therefore, these candidate rSNPs were excluded from further study analysis.

4.2.3 Region 3: Candidate rSNPs rs226198 and rs6880209

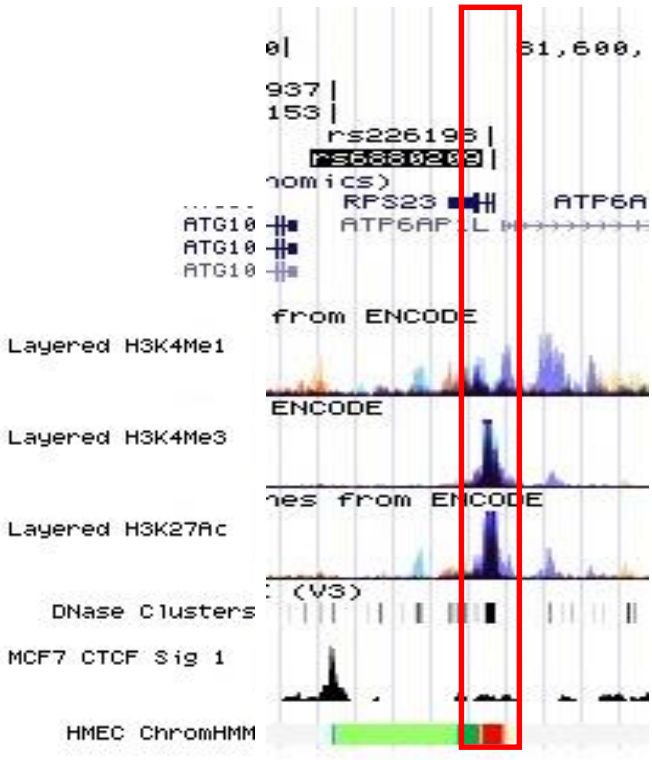


Figure 20: Genomic view of the region 3: rSNPs rs226198 and rs6880209 with functional evidence. From top to bottom of the figure are shown the candidate rSNPs, the RefSeq genes mapped to the area of interest around the 5q14.2 locus, information on histone modifications, DNase I hypersensitivity clusters, CTCF binding region in MCF-7 and chromatin modifications in HMEC according to the Genome Browser (http://genome-euro.ucsc.edu/index.html). These candidate rSNPs overlap active enhancer peak, DHS and CTCF region.

Our *in silico* analysis revealed that both candidate rSNPs rs226198 and rs6880209, in region 3, were found to overlap an active regulatory element peak (H3K27AC, histone H3 acetylated at lysine 27 is a histone that flank active enhancers), as well as an enhancer peak (H3K4Me3, H3 monomethylated at lysine 4 is also a histone mark for enhancer)[37]. Besides these histone modification regions, it was revealed that both were also overlapping a DHS and a CTCF binding region (Figure 20).

As both candidate rSNPs rs226198 and rs6880209 overlap several regulatory element regions, it seemed to be promising candidates. Although rs226198 showed no significance for association with tSNP rs226202 (*RPS23*) with p-value = 0.069, in a DAE mapping analysis consisted of 25 genotyped samples. It did showed a trend for association with the DAE pattern (Figure 21).

For candidate rSNP rs6880209, we verified that this rSNP was associated with the DAE distribution of tSNP rs150934 (*ATP6AP1L*) with p-value = 6.09 x 10^{-3} (Figure 21).

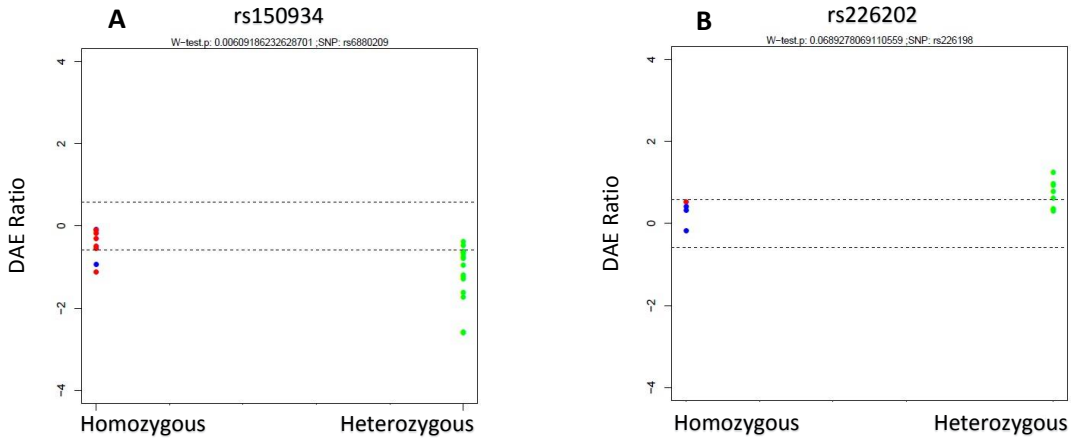


Figure 21: DAE mapping analysis for rs6880209 and rs226198. The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples observed in between the dotted lines have equal expression hence not causing the effect of DAE meanwhile samples observed above or below the dotted lines have unequal expression hence causing the effect of DAE. DAE pattern at the tSNP are identified by grouping the rSNP sample genotypes. [DAE ratio, log 2 (1.5) = 0.584]. (A) rSNP rs6880209 with tSNP rs150934. (B) rSNP rs226198 with tSNP rs226202.

In silico analysis from CCLE database, showed that rs226198 had evidence of c-Myc and POL2 proteins binding, in MCF-7 and MCF10A breast cancer cell lines. In MCF-7 cell line, both proteins have preferential binding to the C allele. For c-Myc, most of the ChIP-seq experiments read counts were higher than 20, with the highest reads count at 565 (Figure 22). Meanwhile, for POL2, most of the ChIP-seq experiments read counts were more than 6 with the highest coverage at 11. In MCF10A, c-Myc and POL2 were also revealed to have a slight preferentially binding to the T allele (Annex 5).

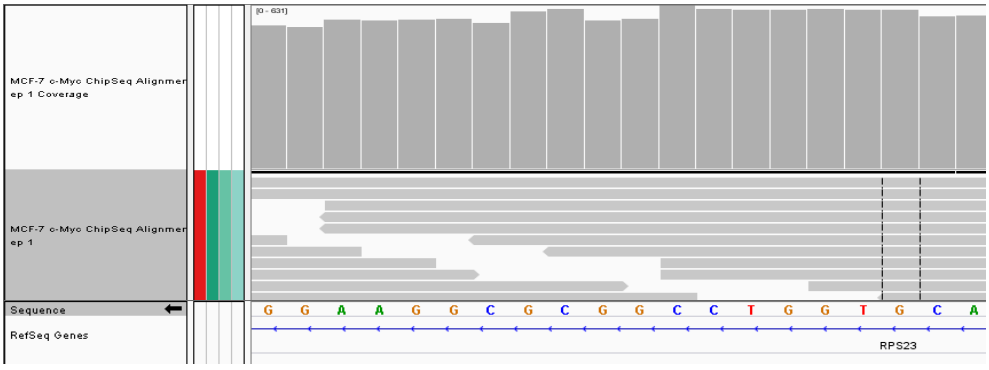


Figure 22: ChIP-seq results in MCF-7 cell line for c-Myc at the rSNP rs226198. Data was taken from CCLE database and viewed with Integrative Genomics Viewer (IGV).

It was also found that for rs6880209, there was c-Myc and POL2 proteins binding both in MCF-7 and MCF10A breast cancer cell lines. In MCF-7, several ChIP-seq experiments showed preferential c-Myc protein binding to the T allele, with the highest ChIP-seq reads count of 188 (83%) (Figure 23). As for the POL2 protein, it also showed preferentially binding to the T allele, with the highest ChIP-seq reads count of 86 (72%). Meanwhile, for the MCF10A cell line, both proteins showed the same binding pattern as for MCF-7 cell line, but c-Myc had ChIP-seq read counts were lower than 20 in this cell type. Overall, all the ChIP-seq experiments point to c-Myc and POL2 proteins having a higher binding affinity for the T allele.

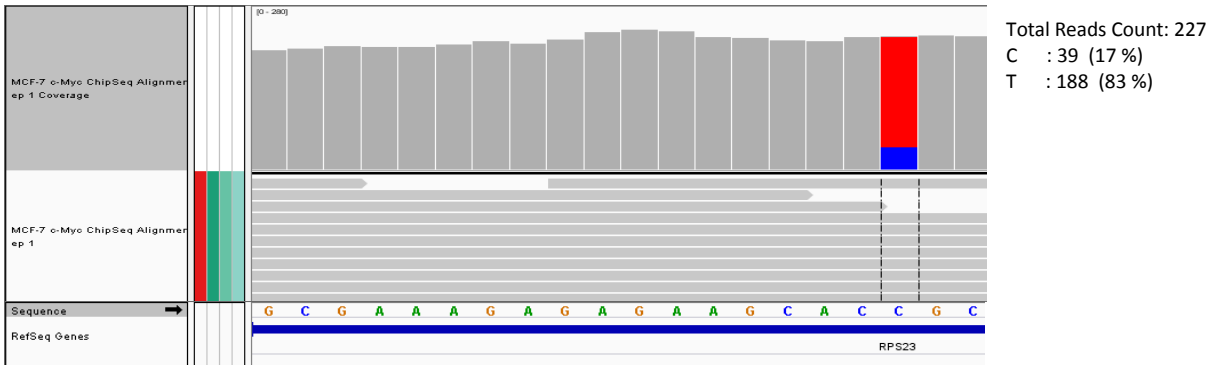


Figure 23: ChIP-seq results in MCF-7 cell line for c-Myc at the rSNP rs6880209. Data was taken from CCLE database and viewed with Integrative Genomics Viewer (IGV).

We decided to include both rs226198 and rs6880209 in further analysis due to evidence of being located in a region with several regulatory elements and strong evidence supporting TF binding. Even though candidate rs226198 showed no significance associated with DAE levels at tSNP rs226202 (*RPS23*), it did showed DAE trending. This rSNP inclusion was justified by the limited number of genotyped samples that might have influenced the DAE mapping analysis.

4.2.4 Region 4: Candidate rSNP rs17247678

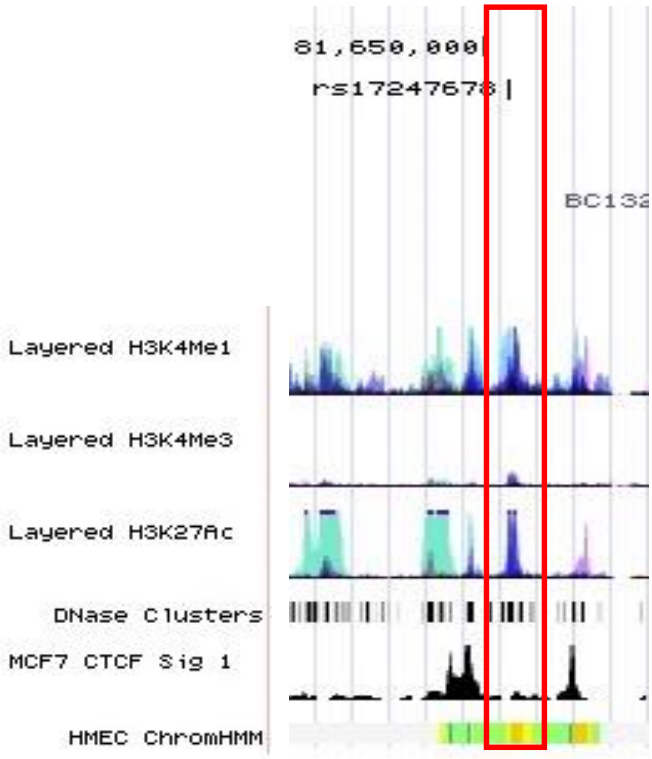


Figure 24: Genomic view of the region 4: rSNP rs17247678 with functional evidence. From top to bottom of the figure is shown the candidate rSNP, the RefSeq gene mapped to the area of interest around the 5q14.2 locus, information on histone modifications, DNase I hypersensitivity site, CTCF binding region in MCF-7 and chromatin modifications in HMEC according to the Genome Browser (http://genome-euro.ucsc.edu/index.html). This candidate rSNP overlaps regulatory elements region such as active enhancer, DHS and CTCF binding.

In region 4, it was found that, besides overlapping enhancer and active regulatory element peaks, candidate rSNP rs17247678 was overlapping a DHS region, as well. DAE mapping analysis performed for this candidate rSNP showed it was associated with the DAE distribution for tSNPs rs150934 (*ATP6AP1L*) with p-value = 2.89 x 10^{-3} and rs7733620 (*ATG10*) with p-value = 2.40 x 10^{-4} (Figure 25). There was no significant association found between this candidate rSNP with tSNP rs226202 (*RPS23*) (p-value = 0.262) (Annex 4).

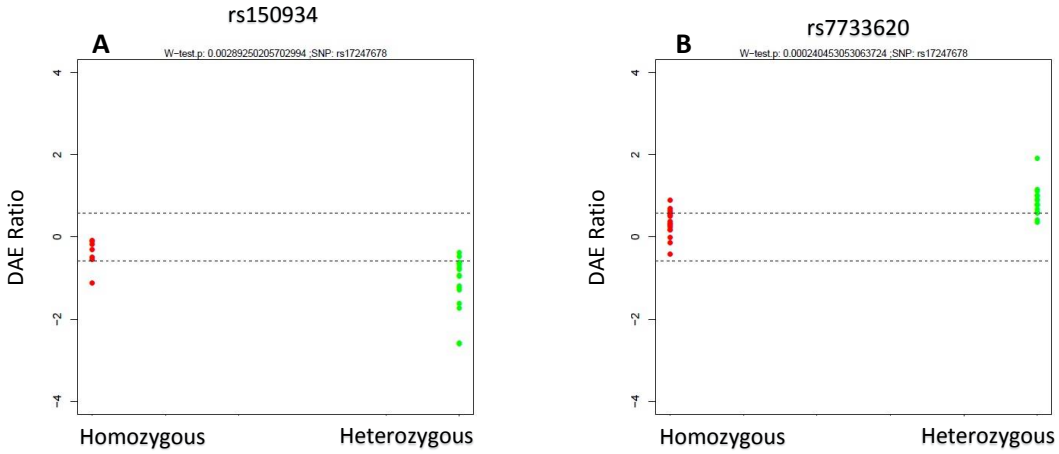


Figure 25: DAE mapping analysis for rSNP rs17247678. The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples observed in between the dotted lines have equal expression hence not causing the effect of DAE meanwhile samples observed above or below the dotted lines have unequal expression hence causing the effect of DAE. DAE pattern at the tSNP are identified by grouping the rSNP sample genotypes. [DAE ratio, log 2 (1.5) = 0.584]. (A) rSNP rs17247678 with tSNP rs150934. (B) rSNP rs17247678 with tSNP rs7733620.

Evidence found in CCLE database showed that there was TF binding to candidate rSNP rs17247678 in MCF10A cell line. STAT3 and c-FOS proteins were found to bind preferentially to A allele. All ChIP-seq experiments had total read counts lower than 20, with the highest ChIP-seq reads count in the coverage region of 28 (Figure 26). Therefore, future ChIP-seq analysis with higher coverage should be conducted in order to have stronger confidence about TF binding.

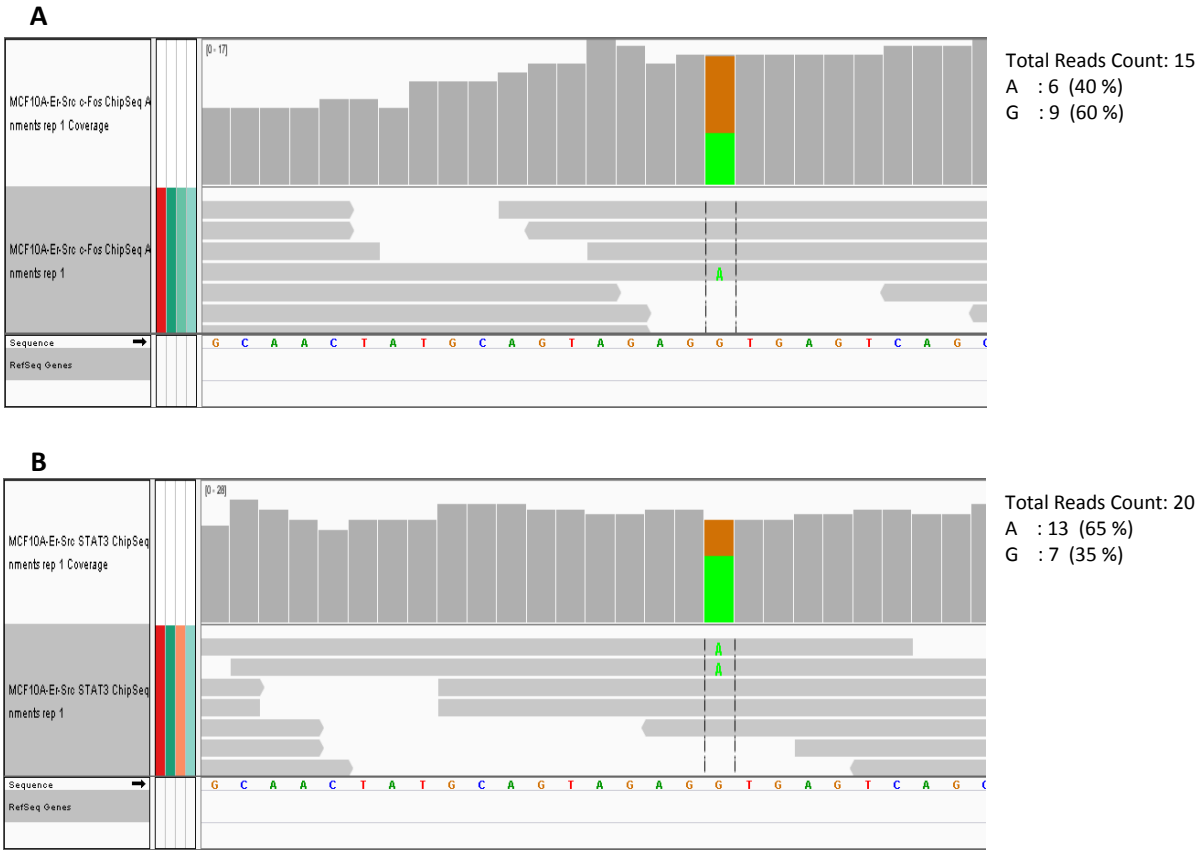


Figure 26: ChIP-seq results in MCF10A cell line for (a) c-FOS and (b) STAT3 at the rSNP rs17247678. Data was taken from CCLE database and viewed with Integrative Genomics Viewer (IGV).

4.3 Expression Quantitative Trait Loci Analysis

In order to assess if total gene expression levels at *ATG10, RPS23* and *ATP6AP1L* genes are associated with the candidate rSNPs (rs226198, rs6880209 and rs17247678) genotypes, we performed eQTL analysis. We had total expression data, from microarrays experiments performed previously for 26 samples for *ATG10* and *ATP6AP1L*. However, we did not have any data for the *RPS23*. Therefore, we conducted gene expression analysis using Taqman real-time PCR for this gene and obtained total expression results for 22 samples (during quality control, five samples without triplicate results were excluded). From eQTL analysis, we found no significant evidence of eQTL in total gene expression with our rSNPs genotypes (Annex 6).

4.4 Linkage Disequilibrium Structure and Haplotypes Block

All five tSNPs, candidate rSNPs with strong functional evidence (rs226198, rs6880209 and rs17247678) and a proxy for the GWAS SNP (rs6888977) were analysed for LD structure and haplotypes using the genotype information of our samples. We did not have genotype information for the GWAS SNP rs7707921, so we used a proxy SNP that was in complete LD (r^2 = 1 in CEU HapMap population), rs6888977, which was genotyped from our previous experiment. For rSNP rs226198, we only had genotype information for 47 out of our 64 samples.

Haplotype analysis revealed that the region where our candidate rSNPs, the GWAS proxy variant and the tSNPs are located, is divided in two haplotype blocks. All the SNPs belong to one of the blocks except for rs17247578. As we can see in Figure 27, Block 1, includes: four tSNPs (rs10068160, rs7733620, rs10044824, rs226202) which corresponded to *ATG10* and *RPS23* genes, proxy SNP for GWAS (rs6888977) in the *ATG10* gene and finally the rSNP rs226198. Block 2, contains one tSNP rs150934 located in the *ATP6AP1L* gene and one rSNP rs6880209. Meanwhile our rSNP rs17247678 was located outside of the haplotype blocks.

Haplotype analysis also revealed that Block 1, had eight haplotypes and Block 2, there had four haplotypes (Figure 28).

The minor allele of the GWAS SNP rs7707921 (T allele) that confers protection to breast cancer risk, is represented Figure 24 by the C allele of rs6888977, which that is present in Haplotypes 2, 5 and 8. Meanwhile from our DAE data, the minor alleles from the tSNPs rs10068160 (G), rs7733620 (T) and rs10044824 (C) from *ATG10* were over expressed than the common alleles (highlighted with red box). The minor allele of the tSNP rs226202 (C) corresponds to less expression of the *RPS23* gene (highlighted with blue box) and minor allele of the tSNP rs150934 (C) correspond to higher expression of the from *ATP6AP1L* gene (highlighted with green box).

Interestingly, Haplotype 2 in Block 1 (frequency approximately 20% includes all the alleles preferentially expressed of the gene *ATG10*, the allele that less expressed of the gene *RPS23,* and the allele that confers protection of the GWAS study. Furthermore, it is mostly inherited together with Haplotype 10 from Block 2, which contains the preferentially expressed allele of the gene *ATP6AP1L*. Interestingly, Haplotype 2 also contains the G allele of rs226198 and Haplotype 10 the T allele of rs6880209, which were observed to have higher binding affinity to c-Myc and POL2. Haplotype 1 (frequency of approximately 50%), has the alternatives alleles of Haplotype 2, including the GWAS risk allele.

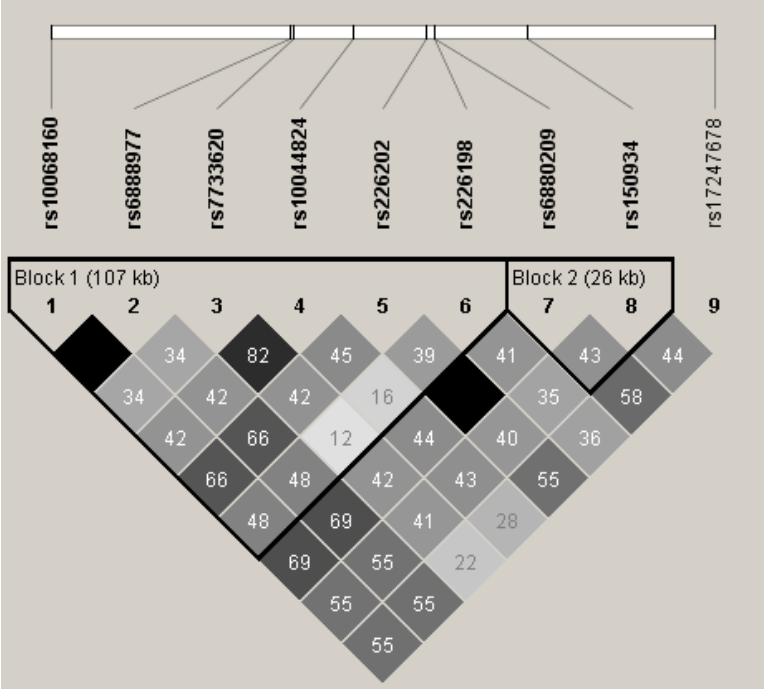


Figure 27: Linkage disequilibrium plot for the analysed proxy GWAS SNP (rs6888977), DAE SNPs (rs10068160, rs7733620, rs10044824, rs226202, rs150934) and rSNPs (rs226198, rs6880209, rs17247678). The SNP ID is displayed along the top of the diagram. This plot was obtained from Haploview using r^2 colour scheme (black indicating r^2=1, with different shades of grey indicating 0 < r^2 < 1). In addition, values in the plot indicate r^2 values for pairwise comparisons between the SNPs. Blocks were defined using the Spine method. Black triangles denote the two haplotypes blocks.

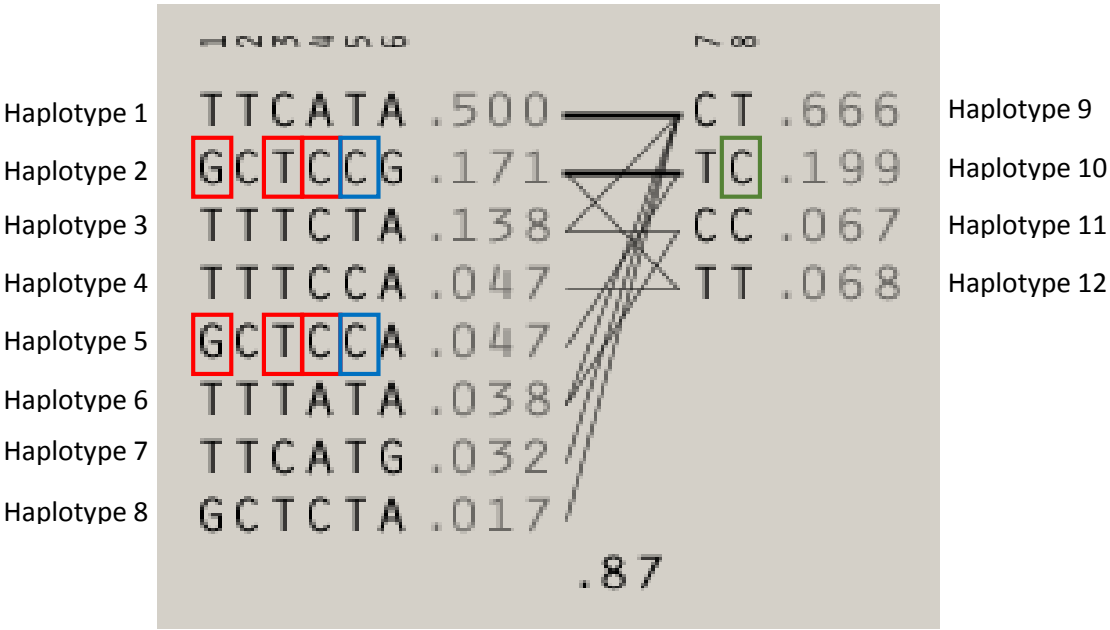


Figure 28: Haplotype block and haplotype frequency in the 5q14.2 locus. The haplotype frequencies are shown to the right side of each haplotype. The SNP numbers at the top correspond to those in the LD plot (Figure 27). The tSNPs used in this analysis are shown in the coloured box. Red correspond to ATG10 gene, blue correspond to RPS23 gene and green correspond to ATP6AP1L gene.

Chapter 5: Discussion

The aim of this study was to identify cis-acting regulatory SNPs that may be responsible for the association identified at the meta-analysis GWASs SNP rs7707921. The locus chosen for analysis in this thesis contained one GWAS SNP and five DAE SNPs. We began by identifying SNPs in LD with GWAS SNP. Out of the 125 initial candidate rSNPs obtained from *in silico* analysis, with an LD cut-off $r^2 \geq 0.4$ with the GWAS SNP, only six were selected to be analysed as it showed evidence of being cis-acting regulatory variants. The cut-off were chosen based on DAE distribution seen in tSNPs showing a trend similar to scenario 2 in Xiao et. al.[22], suggesting a strong but incomplete LD, between the regulatory SNP and the transcribed SNP. Our study identified three potential rSNPs: rs226198, rs68880209 and rs17247678. These might be involved in the regulation of gene(s) showing DAE, and may also be responsible for the risk captured by the GWAS SNP rs7707921.

Our findings revealed that candidate rSNPs rs226198 and rs68880209, in addition to being located in the promoter region shared by the *RPS23* and *ATP6AP1L* genes, also have preferential binding of c-Myc and POL2 (T and C alleles, respectively). These results were observed in two breast cancer cell lines, although for one the number of reads was low ($n \leq 20$). DAE mapping analysis indicated that rs68880209 was associated with DAE levels of the *ATP6AP1L* gene, and rs226198 showed a trend for association with *RPS23* DAE ratios, although the latter did not reach significance. However, it is important to mention that there was not enough genotype information for rs226198, which might hinder the results.

In the case of candidate rSNP rs17247678, we found it is located in a STAT3 binding site, with higher binding affinity to the A allele in MCF10A cell line. Once more, as the number of reads was low ($n \leq 20$), we cannot conclude with confidence on the preferential allelic binding. Nevertheless, it is reassuring that c-FOS, a common co-factor of STAT3, has a binding site located immediately upstream of this position.

We decided to exclude the three other candidate rSNPs based on insufficient evidence supporting their role as regulatory SNPs. Candidate rSNP rs2406909 is located in a regulatory element and DAE mapping analysis revealed it was associated with DAE ratios in tSNPs

rs10044824 and rs7733620 (both in *ATG10* gene*).* However, EMSA experiments showed equal TF binding affinity to both alleles, which cannot explain the DAE pattern obtained for these tSNPs. The band shift seen in the EMSA corresponded to a mass weight ranging from 63 kDa to 75 Da, suggesting the bound protein could have mass weight within that range. Thus, based on the PWM analysis in Table 2, the possible bound protein could either be YY1 or Pou2f2. However, this can only be verified with further analysis by performing competition assays or super shift assays in EMSA[42,43], or by using mass spectrometry[52]. Nevertheless, even though this candidate rSNP could not possibly be the SNP associated with DAE at rs10044824 and rs7733620, it is probably in high LD with the actual rSNP that is regulating these genes.

As for the other two candidate rSNPs, rs10036937 and rs2407153, both were located in a region that has CTCF binding. CTCF binding is known to mediate long-range chromatin looping, which facilitates the activation or insulation of regulatory elements[51]. Nevertheless, both candidate rSNPs had low ChIP-seq reads count ($n \leq 20$), which poses uncertainties on the difference of binding affinity between alleles seen in the experiments. Nothnagel et. al. suggested reads count should have higher value of 20, in order to infer allelic imbalance binding [53]. DAE mapping analysis showed that only rSNP rs2407153 had a significant associated at tSNP rs10044824. However, it is important to mention that there was not enough genotype information for rs2407153, which might hinder the results. Therefore, it supported the idea that these two SNPs were not the cis-acting regulatory SNP responsible for DAE at these genes.

Undoubtedly, there were more candidate rSNPs in the preliminary list than the six SNPs described in this work. However, filtering criteria were applied in order to rank potential rSNPs on priority to be studied. These criteria were: (1) evidence of being located in regulatory elements, (2) located in a DHS, (3) preferentially allelic TF binding, (4) potential of being involved in CTCF binding. This approach shortlists candidate rSNPs but there is the possibility that some SNPs with a functional effect have been missed (false negatives). Currently we believe that the regulation of the genes in locus 5q14.2 is a complex interaction between several rSNPs, and we are carrying further experiments to confirm this.

We have taken two approaches in order to identify cis-acting regulatory variants: one is through eQTL search and the other is based on DAE mapping[23]. Therefore to compare these two

approaches, we also performed eQTL analysis with our samples in order to assess total gene expression association with our rSNPs genotypes. For eQTL analysis, we used total gene expression data from 26 samples for *ATG10* and *ATP6AP1L* and 22 samples for *RPS23*.

In Michailidou et. al. study, they found a strong association of total expression in *RPS23* with GWAS SNP rs7707921 (with AK092335_1_1696 probe, $p = 2.0 \times 10^{-15}$ in tumours, $p = 1.11 \times 10^{-4}$ in normal tissue, with AK092335_1_1335, $p = 4.30 \times 10^{-11}$ in tumours and $p = 3.23 \times 10^{-2}$ in normal tissue) and a weak association with the expression of *ATP6AP1L* (p = 5.6 x 10-5 in tumours and p = 0.066 in normal tissue)[12]. However, they were not able to identify a causal gene.

Through our DAE approach, we were able to identify cis-regulation in the three genes in locus 5q14 (*ATG10, RPS23* and *ATP6AP1L*) and identified three possible rSNPs that might be involved in the regulation of these genes. We performed eQTL analysis with our rSNP genotypes information, and found that there was no significant association between the rSNPs and the three genes, supporting our hypothesis that DAE is a more efficient approach to identify cis-acting regulatory variants.

One possible explanation for eQTL to fail to detect cis-regulating signals has to do with feedback control mechanisms in the cell. Take for example, three genotypes, AA, AB and BB. The expression of allele A is higher but the overall expression of the homozygous AA is strongly controlled by the feedback control mechanism. Meanwhile homozygous BB has lower expression, therefore feedback control mechanism is not activated. As for heterozygous AB, it has intermediate levels of feedback control. Consequently, in total gene expression, variations between individuals are not strongly evident due to this feedback control mechanism. However, allelic output in the heterozygous AB, still indicate the difference between cis-acting regulatory alleles. Even if a strong feedback control is applied to heterozygous AB, it still can be measured using DAE, as both alleles in heterozygous are under equal feedback control[54].

A housekeeping gene was used as internal control for gene expression. Housekeeping genes are recognised as cellular maintenance gene that regulate basic cellular function. Therefore, we used *GAPDH* gene to normalise the *RPS23* gene expression[55]. Thus, ensuring that the results obtained were not influence by individual genetic makeup. It is also important to mention that we performed the eQTL analysis with a low number of samples and most of the

rSNPs did not have all three genotypes. Thus, the low number of samples have an effect on the statistical power of the analysis. Nevertheless, as we have identified three potential rSNPs supported by strong evidence through DAE and eQTL approaches, further analysis should be conducted in order to understand the biological mechanisms involved in expression of the corresponding genes (*ATG10, RPS23* and *ATP6AP1L*) and their relation to breast cancer.

c-Myc is a TF that integrates the cell cycle machinery with cell adhesion, cellular metabolism and apoptotic pathways[56]. c-Myc deregulation including gene amplification, transcriptional regulation, mRNA and protein stabilisation that correlate to loss of tumour suppressors and activation of oncogenic pathways, contributes to breast cancer development, progression and associated with poor outcomes[57,58].

POL2 is known as a component of the general transcriptional mechanism and mediates the transcription process in cells. One study used POL2 binding data in order to predict translational regulation of anti-oestrogen resistance in breast cancer[59].

STAT3 is an important mediator of the transcriptional changes during tumorigenesis process and has been implicated in many types of cancers including breast cancer[60].

c-FOS has been reported to differently activate transcription of target genes and induce morphological changes in carcinoma[61].

Higher or lower binding affinity of these TFs could influence gene expression levels in an allele specific way, as observed with DAE ratios of the three genes in our study.

Further analysis such as 3C (chromatin conformation capture) allows verification of interaction between one target site with another target site[62–64]. For example, the enhancer downstream of *ATP6AP1L* is contributing to its expression regulation, would it be able to come in contact with the promoter of *ATP6AP1L*. 4C (circularised chromatin conformation capture) allows one selected site to be used as a marker and the genome is screened for sequences that is in contact with the selected site[65]. As for 5C (carbon-copy chromatin conformation capture), it allows concurrent determination of interactions between multiple sequences[66].

There is very limited information on these three genes in the literature, particularly in relation to cancer. *ATG10* gene is involved in autophagy and the ATG family genes are key regulators of this process. One study has reported an association between increased *ATG10*

expression and tissue invasion and lymphovascular metastasis, in colorectal cancer[67,68]. This is in opposition with our results, in which we see an association between higher expression and protection for breast cancer.

RPS23 encodes the ribosomal protein S23, a component of the 40S subunit. This protein belongs to the S12P family of ribosomal proteins, is located in the cytoplasm and is crucial in the development of human diseases. A study has reported over expression of *RPS23* in early and advanced stages of colorectal adenocarcinomas[69], supporting a possible role of this gene as an oncogene that is in accordance with what we observe in our study.

ATP6AP1L also known as vacuolar H+-ATPase, is involved in controlling intracellular compartments in eukaryotic cells including intracellular membrane transport, pro-hormone processing and transport of neurotransmitters[70]. Several studies revealed an association of high activity of vacuolar H+-ATPase with breast cancer and regulation of vacuolar H+-ATPase as potential for cancer therapy[71–73]. However, this is in opposition with our results, in which we see an association between higher expression and protection for breast cancer.

Although several studies have investigated expression levels for these three genes in different types of cancer, their role in breast cancer predisposition and development has not been elucidated yet. Therefore, future studies on these genes might be of interest, in order to better understand the biological effects they may have in signalling pathways related to breast cancer.

From our observation, we saw the possibility of a complex network of rSNPs regulating several genes in the 5q14.2 locus, which causes the signal in the GWAS SNP rs7707921 study. However, for now, we do not have the data to test these SNPs association with breast cancer risk.

Our haplotypes analysis showed that combination of rSNPs (rs226198, rs6880209 and rs17247678) could possibly be regulating three genes (*ATG10, RPS23* and *ATP6AP1L*) and lead to risk or protection to breast cancer.

From our haplotype analysis, combination of Haplotype 1, which contained proxy GWAS SNP rs6888977 and rSNP rs226198 with Haplotype 9 that contained rSNP rs6880209, infer risk to breast cancer. This Haplotype 1 and Haplotype 9 combination has tSNPs rs10068160, rs7733620,

rs10044824 and rs226202 in Haplotype 1 and tSNP rs150934 in Haplotype 9. That correspond to over expression of *RPS23* and under expression of *ATG10* and *ATP6AP1L*. Contrary to Haplotype 1 and Haplotype 9 combination, Haplotype 2 and Haplotype 10 combination infers protection to breast cancer. However, in order to verify this hypothesis, further association analysis on both Haplotype 1/Haplotype 9 and Haplotype2/Haplotype 10 combination with breast cancer should be conducted.

Chapter 6: Conclusion

From our study, we were able to identify possible rSNPs that could be the causal variants of the risk-associated SNP rs7707921 for breast cancer. Three rSNPs (rs226198, rs6880209 and rs17247678) were identified as cis-acting regulators that seem to regulate three genes (*ATG10, RPS23* and *ATP6AP1L*) in the 5q14.2 locus. Here, we propose model in which c-Myc and POL2 binding to the common allele of rs226198 (A) and rs6880209 (C) lead to over expression of *RPS23* and under expression of *ATG10*, respectively. Meanwhile, STAT3 binding to the common allele of rs17247678 (G) and c-FOS binding to its vicinity lead to under expression of *ATP6AP1L* (Figure 29). Risk to breast cancer could possibly be due to complex network of rSNPs regulation all three genes. Therefore, further analysis is needed to understand the underlying mechanism involved between the regulatory variants and the genes.

Subsequently, these findings could contribute to better understanding of the biology for breast cancer as well as assist in the future development of cancer prevention and treatment therapies.

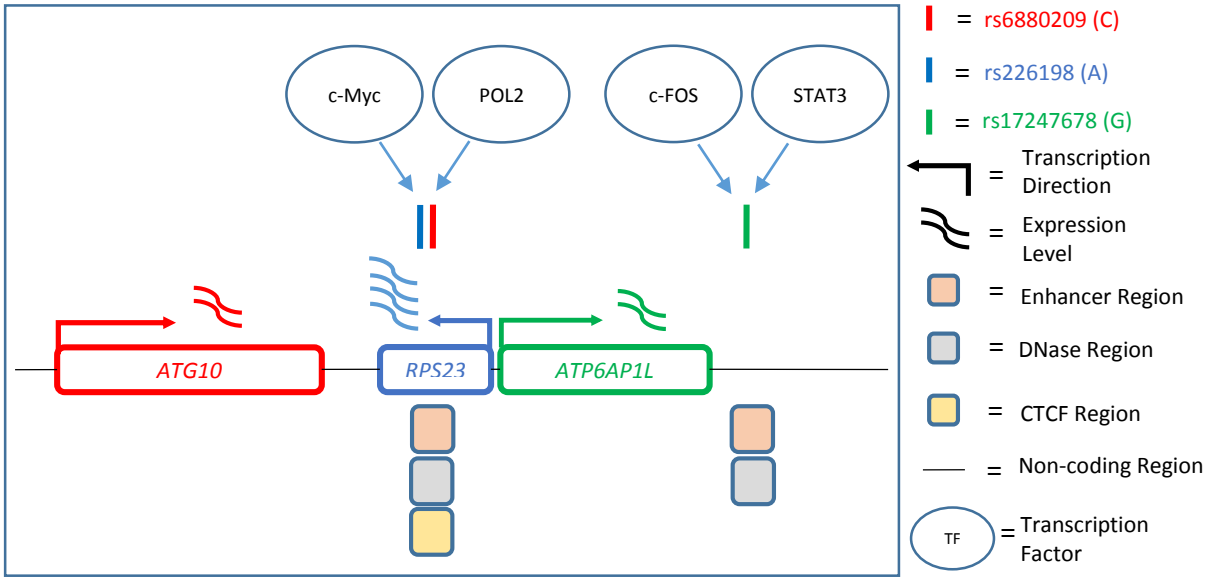


Figure 29: Scheme representing the putative gene-regulation mechanism of the candidate rSNPs identified in this thesis. This putative gene-regulation mechanism confers risk to breast cancer associated with GWAS SNP rs7707921. c-Myc and POL2 binding to frequent allele rs226198 (A) and rs6880209 (C) lead to over expression of RPS23 and under expression of ATG10, respectively. Meanwhile STAT3 and c-FOS binding to frequent allele rs17247678 (G) and lead to under expression of ATP6AP1L.

Bibliography

1. Hanahan, D. & Weinberg, R. a. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).

2. Hanahan, D. & Weinberg, R. a. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).

3. Shen, Z. Genomic instability and cancer: An introduction. *Journal of Molecular Cell Biology* **3,** 1–3 (2011).

4. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11,** 220–228 (2010).

5. Ferlay, J. *et al.* Cancer incidence and mortality worldwide : Sources , methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136,** 359–386 (2015).

6. McPherson, K., Steel, C. M. & Dixon, J. M. ABC of breast diseases. Breast cancer--epidemiology, risk factors and genetics. *BMJ Br. Med. J.* **309,** 1003–1006 (1994).

7. Yang, X. R. *et al.* Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the breast cancer association consortium studies. *J. Natl. Cancer Inst.* **103,** 250–263 (2011).

8. Maxwell, K. & Domchek, S. Familial Breast Cancer Risk. *Curr. Breast Cancer Rep.* **5,** 170–182 (2013).

9. Economopoulou, P., Dimitriadis, G. & Psyrri, a. Beyond BRCA: New hereditary breast cancer susceptibility genes. *Cancer Treat Rev* **41,** 1–8 (2015).

10. Ghoussaini, M., Pharoah, P. D. P. & Easton, D. F. Inherited Genetic Susceptibility to Breast Cancer. *Am. J. Pathol.* **183,** 1038–1051 (2013).

11. Antoniou, a *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* **72,** 1117–1130 (2003).

12. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47,** 373–380 (2015).

13. Buckland, P. R. The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1762,** 17–28 (2006).

14. Chorley, B. N. *et al.* Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. *Mutat. Res. - Rev. Mutat. Res.* **659,** 147–157 (2008).

15. Wang, X., Tomso, D. J., Liu, X. & Bell, D. a. Single nucleotide polymorphism in transcriptional regulatory regions and expression of environmentally responsive genes. *Toxicol. Appl. Pharmacol.* **207,** 84–90 (2005).

16. Edwards, S. L., Beesley, J., French, J. D. & Dunning, M. Beyond GWASs: Illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93,** 779–797 (2013).

17. Hunter, D. J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39,** 870–874 (2012).

18. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447,** 1087–1093 (2007).

19. Ahmed, S. *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* **41,** 585–590 (2009).

20. Thomas, G. *et al.* A multi-stage genome-wide association in breast cancer identifies two novel risk alleles at 1p.11.2 and 14q24.1 (RAD51l1). *Nature* **41,** 579–584 (2010).

21. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41,** 324–328 (2009).

22. Xiao, R. & Scott, L. J. Detection of cis-acting regulatory SNPs using allelic expression data. *Genet. Epidemiol.* **35,** 515–525 (2011).

23. Pastinen, T., Ge, B. & Hudson, T. J. Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet.* **15,** 9–16 (2006).

24. Serre, D. *et al.* Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic Cis-acting mechanisms regulating gene expression. *PLoS Genet.* **4,** 1–16 (2008).

25. Fletcher, O. *et al.* Novel breast cancer susceptibility locus at 9q31.2: Results: of a genome-wide association study. *J. Natl. Cancer Inst.* **103,** 425–435 (2011).

26. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45,** 353–361 (2013).

27. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).

28. Ward, L. D. & Kellis, M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** 1–5 (2012).

29. Chen, X., Guo, L. & Fan, Z. Learning Position Weight Matrices from Sequence and Expression Data. *Comput Syst Bioinform Conf.* **6,** 249–260 (2007).

30. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).

31. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210 (2002).

32. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* **41,** 991–995 (2013).

33. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

34. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (2010).

35. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347,** 394–404 (2015).

36. Kent, W. J. *et al.* The Human Genome Browser at UCSC The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

37. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–86 (2014).

38. Stransky, N. *et al.* Integrative analysis of genomic and pharmacologic data from the Cancer Cell Line Encyclopedia. *Cancer Res.* **70,** 105 (2010).

39. Robinson, James T.; Thorvaldsdóttir, Helga; Winckler, Wendy; Guttman, Mitchell; Lander, Eric S.; Getz, Gad; Mesirov, J. P. Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

40. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).

41. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).

42. Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* **2,** 1849–1861 (2007).

43. Holden, N. S. & Tacon, C. E. Principles and problems of the electrophoretic mobility shift assay. *J. Pharmacol. Toxicol. Methods* **63,** 7–14 (2011).

44. Buratowski, S. & Chodosh, L. a. Mobility shift DNA-binding assay using gel electrophoresis. *Curr. Protoc. Mol. Biol.* **Chapter 12,** Unit 12.2 (2001).

45. Meyer, K. B. *et al.* Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol.* **6,** 1098–1103 (2008).

46. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).

47. Heid Stevens, J., Livak, K.J., and Williams, P.M., C. a. Real time quantitative PCR. *Genome Res* **6,** 986–994 (1996).

48. Ahmed, F. E., Oncology, R. & Jenkins, L. W. Quantitative Real-time RT-PCR : Application to Carcinogenesis. **332,** 317–332 (2005).

49. Cockerill, P. N. Structure and function of active chromatin and DNase i hypersensitive sites. *FEBS J.* **278,** 2182–2210 (2011).

50. Madrigal, P. & Krajewski, P. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front. Genet.* **3,** 1–3 (2012).

51. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17,** 520–527 (2001).

52. Lin, D., Tabb, D. L. & Yates, J. R. Large-scale protein identification using mass spectrometry. *Biochim. Biophys. Acta - Proteins Proteomics* **1646,** 1–10 (2003).

53. Nothnagel, M. *et al.* Statistical inference of allelic imbalance from transcriptome data. *Hum. Mutat.* **32,** 98–106 (2011).

54. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* **11,** 533–538 (2010).

55. Turabelidze, A., Guo, S. & Dipietro, L. a. Importance of housekeeping gene selection for accurate reverse transcription-quantitative polymerase chain reaction in a wound healing model. *Wound Repair Regen.* **18,** 460–466 (2010).

56. Dang, C. V *et al.* Function of the c-Myc oncogenic transcription factor. *Exp. Cell Res.* **253,** 63–77 (1999).

57. Xu, J., Chen, Y. & Olopade, O. I. MYC and Breast Cancer. *Genes Cancer* **1,** 629–640 (2010).

58. Liao, D. J. & Dickson, R. B. c-Myc in breast cancer. *Endocr. Relat. Cancer* **7,** 143–164 (2000).

59. Zhang, D., Wang, G. & Wang, Y. Transcriptional regulation prediction of antiestrogen resistance in breast cancer based on RNA polymerase II binding data. *BMC Bioinformatics* **15 Suppl 2,** S10 (2014).

60. Fleming, J. D. *et al.* STAT3 acts through pre-existing nucleosome- depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer. *Epigenetics Chromatin* **8,** 1–14 (2015).

61. Andersen, H. *et al.* The ability of Fos family members to produce phenotypic changes in epithelioid cells is not directly linked to their transactivation potentials. *Oncogene* **21,** 4843–4848 (2002).

62. De Wit, E. & de Laat, W. A decade of 3C technologies: Insights into nuclear organization. *Genes Dev.* **26,** 11–24 (2012).

63. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311 (2002).

64. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat. Methods* **3,** 17–21 (2006).

65. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38,** 1341–1347 (2006).

66. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).

67. Qin, Z. *et al.* Potentially functional polymorphisms in ATG10 are associated with risk of breast cancer in a Chinese population. *Gene* **527,** 491–495 (2013).

68. Jo, Y. K. *et al.* Increased Expression of ATG10 in Colorectal Cancer Is Associated with Lymphovascular Invasion and Lymph Node Metastasis. *PLoS One* **7,** 8–13 (2012).

69. Lau, T. P. *et al.* Pair-wise comparison analysis of differential expression of mRNAs in early and advanced stage primary colorectal adenocarcinomas. *BMJ Open* **4,** 1–11 (2014).

70. Nishi, T. & Forgac, M. The vacuolar (H+)-ATPases--nature's most versatile proton pumps. *Nat. Rev. Mol. Cell Biol.* **3,** 94–103 (2002).

71. Capecci, J. & Forgac, M. The function of vacuolar ATPase (V-ATPase) a subunit isoforms in invasiveness of MCF10a and MCF10CA1a human breast cancer cells. *J. Biol. Chem.* **288,** 32731–32741 (2013).

72. Sennoune, S. R. *et al.* Vacuolar H+-ATPase in human breast cancer cells with distinct metastatic potential: distribution and functional activity. *Am. J. Physiol. Cell Physiol.* **286,** C1443–C1452 (2004).

73. Hernandez. Intracellular Proton Pumps as Targets in Chemotherapy: V-ATPases and Cancer. *Curr. Pharm. Des.* **18,** 1383–1394 (2012).

Annex

Annex 1: Score assigned by RegulomeDB according to the functional evidence.

Table containing the SNP scores assigned by RegulomeDB according to the functional evidence in the database.

Score	Supporting Data
Likely to affect binding and linked to expression of a gene target	
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
Likely to affect binding	
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
Less likely to affect binding	
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
Minimal binding evidence	
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

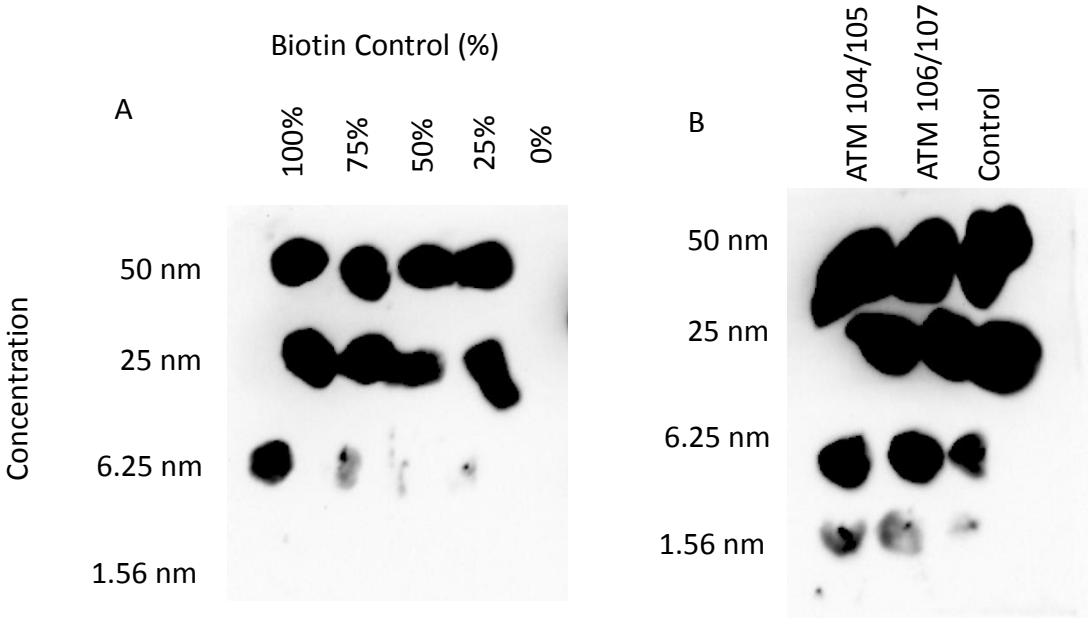
Annex 2: Oligonucleotide sequences designed for PCR and EMSA.

In the table are represented the three selected SNPs to analyse. ATM 100 – ATM 103 are oligonucleotide sequences for PCR analysis meanwhile ATM 104 – ATM 107 are oligonucleotide sequences for EMSA analysis. Minor allele is shown first and the common allele is shown second. This table also show the sequences designed for both alleles for each SNP.

Primer Name	SNP	Sequence
ATM 100	rs2407153 (Forward)	GCCCTTGCTTTAACTTTTGTATTG
ATM 101	rs2407153 (Reverse)	GATTCTTCTCCAGGGCTTCGTGAT
ATM 102	rs226198 (Forward)	AGGATGGGTGAGCTGTTGTG
ATM 103	rs226198 (Reverse)	CGGAGCTTCCTAGCAGTACG
ATM 104	rs2406909 : T allele (Sense)	ATTCCTCGGTTTGATTTCCATATTCCTAGTGTAAGT
ATM 105	rs2406909 : T allele (Anti Sense)	ACTTACACTAGGAATATGGAAATCAAACCGAGGAAT
ATM 106	rs2406909 : C (Sense)	ATTCCTCGGTTTGATTTCCACATTCCTAGTGTAAGT
ATM 107	rs2406909 : C (Anti Sense)	ACTTACACTAGGAATGTGGAAATCAAACCGAGGAAT

Annex 3: Labelling efficiency for Biotin Control DNA and annealed oligonucleotides.

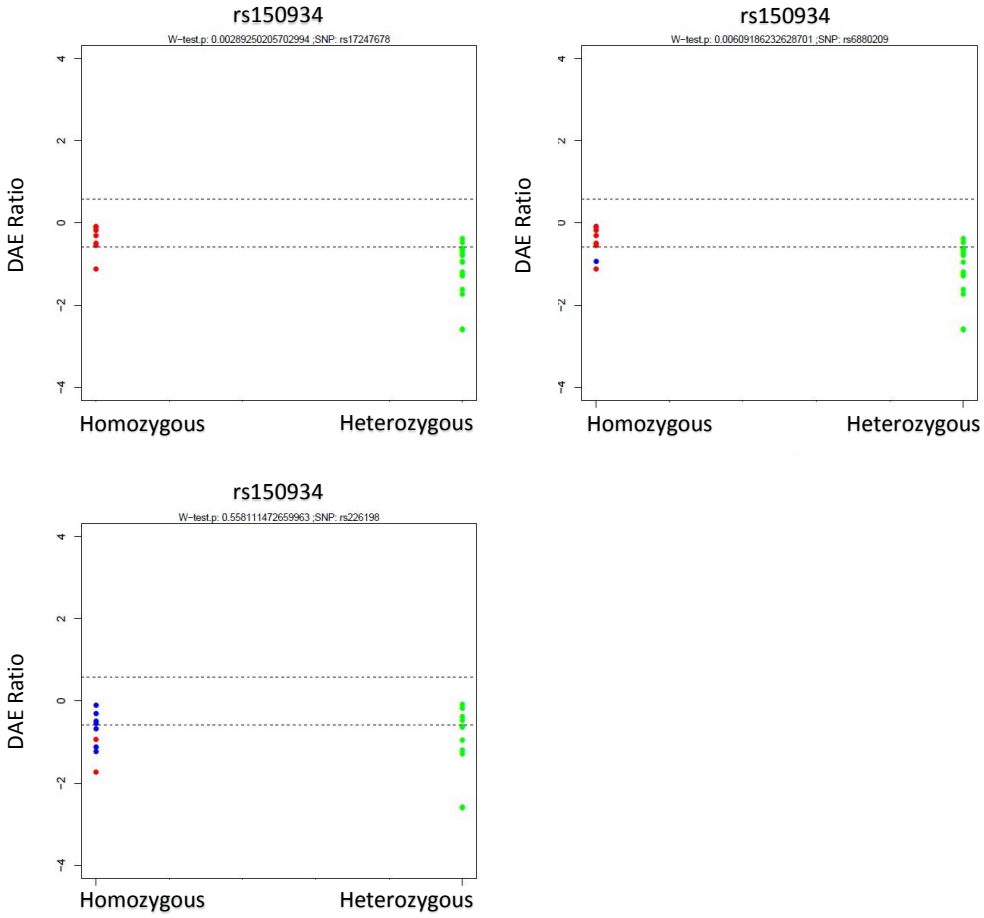
The figure below includes the dilution for each of the standard of the Procedures for Estimating Labelling Efficiency (A), as well as the labelling control of the kit and two different test oligonucleotides (B). ATM 104/105 correspond rs240609 T allele and ATM 106/107 correspond to rs240609 C allele.



Annex 4: DAE Mapping Analysis for tSNPs.

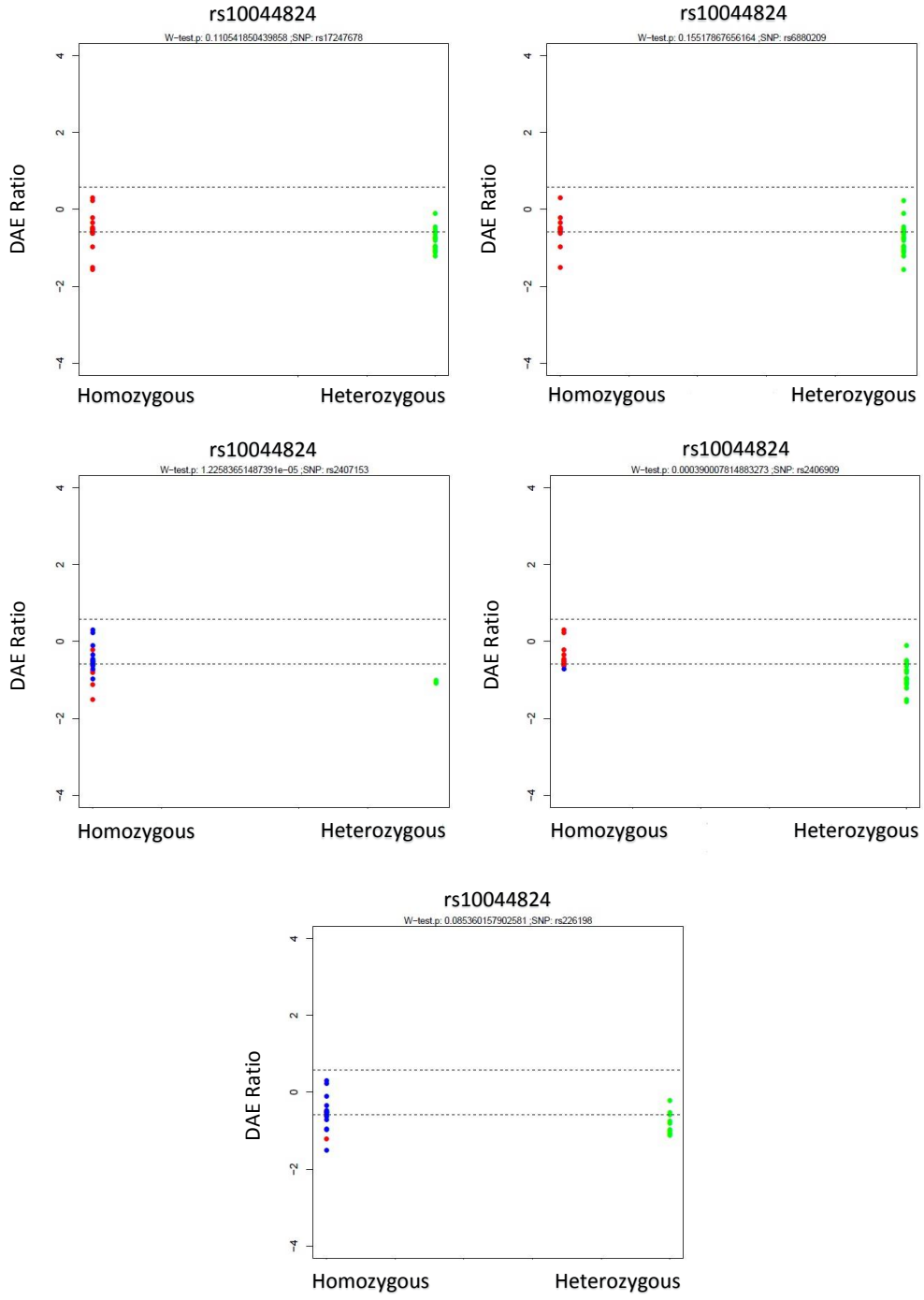
DAE mapping analysis for *ATP6AP1L* at rs150934.

The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples that are homozygous at the rSNP were observed to be located inside the dotted line as tSNP was expressed equally, hence, not causing the DAE effect. Meanwhile samples that are heterozygous at the rSNP were observed to be located above or below the dotted lines as tSNP was expressed unequally, hence, causing the DAE effect. [DAE ratio, log 2 (1.5) = 0.584].



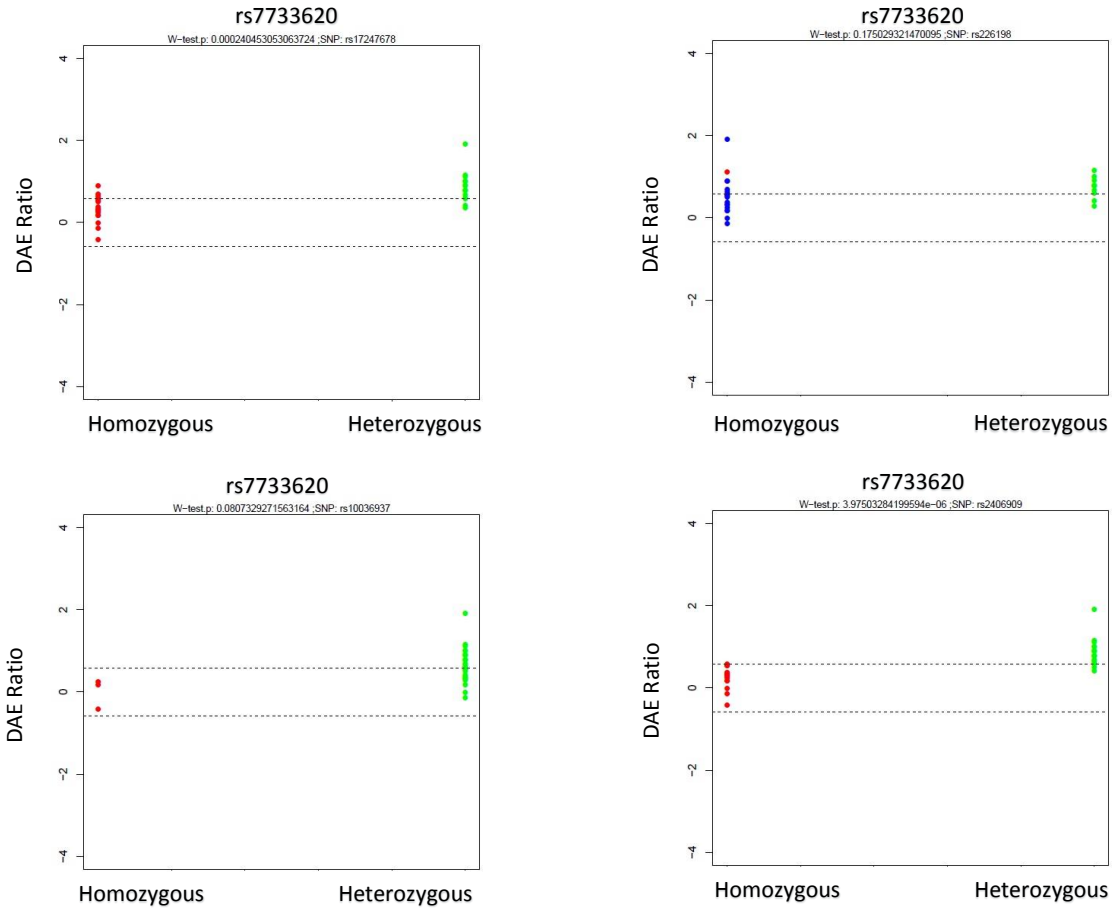
DAE mapping analysis for *ATG10* at rs10044824.

The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples that are homozygous at the rSNP were observed to be located inside the dotted line as tSNP was expressed equally, hence, not causing the DAE effect. Meanwhile samples that are heterozygous at the rSNP were observed to be located above or below the dotted lines as tSNP was expressed unequally, hence, causing the DAE effect. [DAE ratio, log 2 (1.5) = 0.584].



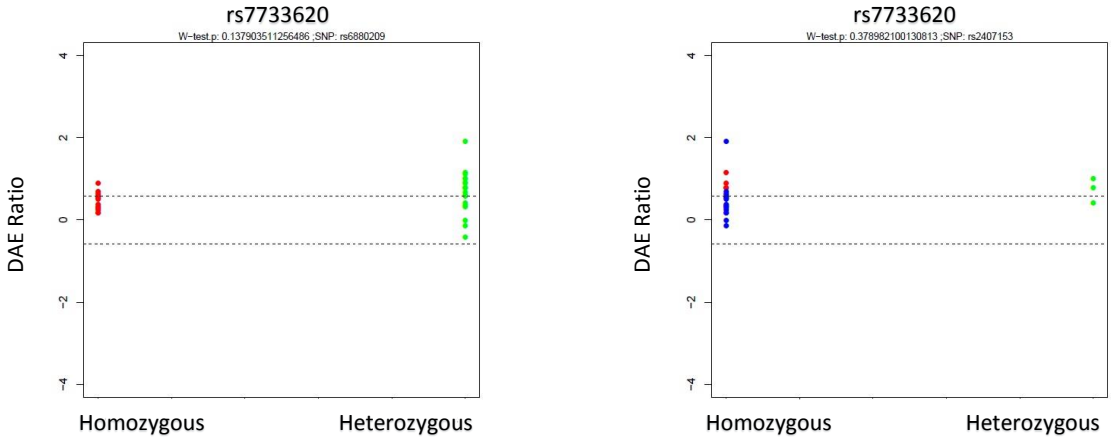
DAE mapping analysis for *ATG10* at rs7733620.

The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples that are homozygous at the rSNP were observed to be located inside the dotted line as tSNP was expressed equally, hence, not causing the DAE effect. Meanwhile samples that are heterozygous at the rSNP were observed to be located above or below the dotted lines as tSNP was expressed unequally, hence, causing the DAE effect. [DAE ratio, log 2 (1.5) = 0.584].



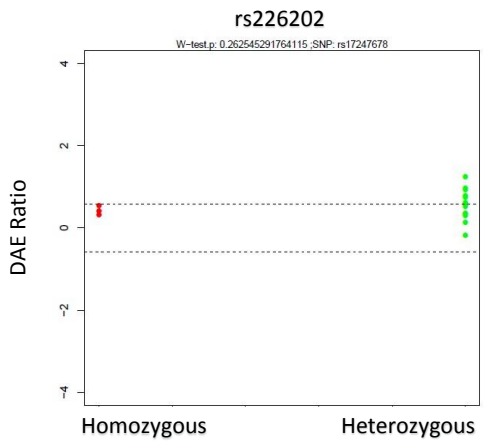
DAE mapping analysis for *ATG10* at rs7733620.

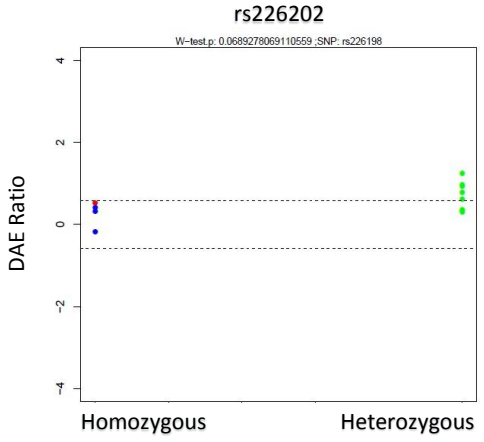
The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples that are homozygous at the rSNP were observed to be located inside the dotted line as tSNP was expressed equally, hence, not causing the DAE effect. Meanwhile samples that are heterozygous at the rSNP were observed to be located above or below the dotted lines as tSNP was expressed unequally, hence, causing the DAE effect. [DAE ratio, log 2 (1.5) = 0.584].

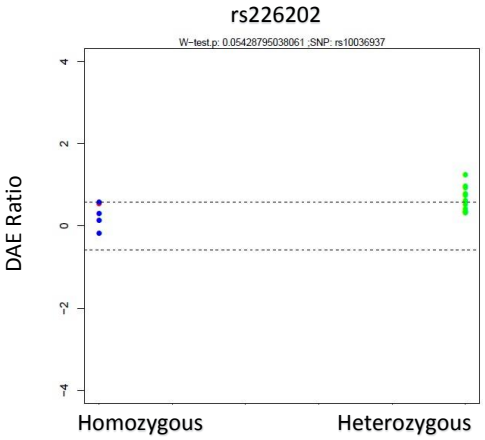


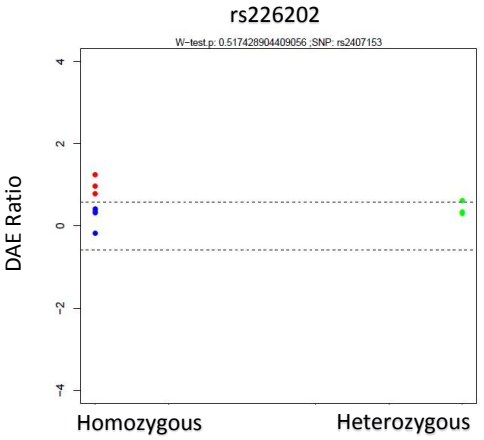
DAE mapping analysis for *RPS23* at rs226202.

The x-axis represent the genotypes meanwhile the y-axis represent the DAE ratio seen at the tSNP. Samples that are homozygous at the rSNP were observed to be located inside the dotted line as tSNP was expressed equally, hence, not causing the DAE effect. Meanwhile samples that are heterozygous at the rSNP were observed to be located above or below the dotted lines as tSNP was expressed unequally, hence, causing the DAE effect. [DAE ratio, log 2 (1.5) = 0.584].









Annex 5: ChIP-seq data for candidate rSNPs in the 5q14.2 locus.

ChIP-seq data retrieved from CCLE database. In the table, the candidate rSNP identification, type of cell lines used in the experiments, total reads count and reads count per allele with their respective percentage are also included.

Candidate rSNP	Cell Line	Protein	Total Reads	Reads	%	Reads	%
				C Allele		G Allele	
rs10036937 (Region 2)	MCF-7	CTCF	4	4	100%	0	0%
			17	17	100%	0	0%
			8	8	100%	0	0%
			1	1	100%	0	0%
			2	2	100%	0	0%
			11	11	100%	0	0%
			24	23	96%	1	4%
			11	11	100%	0	0%
			11	11	100%	0	0%
			12	12	100%	0	0%
			19	19	100%	0	0%
			6	6	100%	0	0%
			11	11	100%	0	0%
	HMF	CTCF	7	5	71%	2	29%
			23	14	61%	9	39%
	HMEC	CTCF	7	4	57%	3	43%
			3	2	67%	1	33%
			8	3	38%	5	63%
			7	4	57%	3	43%

Candidate rSNP	Cell Line	Protein	Total Reads	Reads	%	Reads	%
				G Allele		T Allele	
rs2407153 (Region 2)	MCF-7	CTCF	1	0	0%	1	100%
			2	0	0%	2	100%
			2	0	0%	2	100%
			2	0	0%	2	100%
			1	0	0%	1	100%

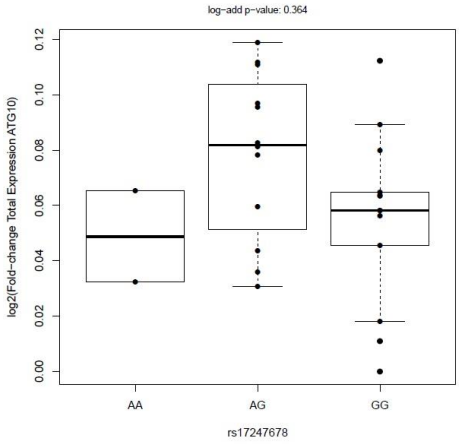
Candidate rSNP	Cell Line	Protein	Total Reads	Reads	%	Reads	%
				C Allele		T Allele	
rs226198 (Region 3)	MCF-7	c-Myc	29	22	76%	7	24%
			20	18	90%	2	10%
			612	565	92%	42	7%
			104	89	86%	15	14%
			30	20	67%	10	33%
			20	17	85%	3	15%
			78	68	87%	10	13%
			38	30	79%	8	21%
		POL2	2	1	50%	1	50%
			5	5	100%	0	0%
			18	11	61%	7	39%
			7	6	86%	1	14%
			7	6	86%	1	14%
			10	6	60%	4	40%
	MCF10A	c-Myc	35	23	66%	12	34%
			16	10	63%	6	38%
			5	2	40%	3	60%
			1	0	0%	1	100%
		POL2	17	8	47%	9	53%
			28	13	46%	15	54%
			18	10	56%	8	44%
			34	17	50%	17	50%

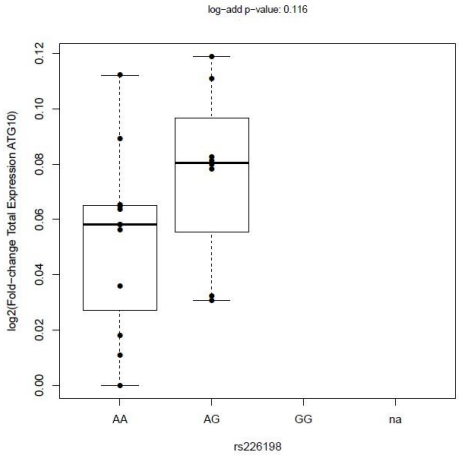
Candidate rSNP	Cell Line	Protein	Total Reads	Reads	%	Reads	%
				C Allele		T Allele	
rs6880209 (Region 3)	MCF-7	c-Myc	41	13	32%	27	66%
			15	6	40%	9	60%
			227	39	17%	188	83%
			71	16	23%	55	77%
			11	3	27%	8	73%
			18	5	28%	13	72%
			58	8	14%	50	86%
			32	3	9%	29	91%
		POL2	18	6	33%	12	67%
			39	14	36%	25	64%
			119	33	28%	86	72%
			105	23	22%	82	78%
			42	12	29%	30	71%
			40	13	33%	27	68%
	MCF10A	c-Myc	12	2	17%	10	83%
			19	6	32%	13	68%
			8	5	63%	3	38%
			8	1	13%	7	88%
		POL2	42	21	50%	21	50%
			63	30	48%	33	52%
			41	18	44%	23	56%
			58	29	50%	29	50%

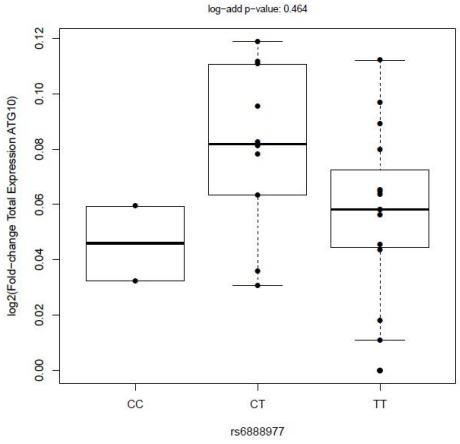
Candidate rSNP	Cell Line	Protein	Total Reads	Reads	%	Reads	%
				G Allele		A Allele	
rs17247678 (Region 4)	MCF10A	c-FOS	15	9	60%	6	40%
			3	1	33%	2	67%
			14	8	57%	6	43%
			7	3	43%	4	57%
			9	4	44%	5	56%
			5	1	20%	4	80%
			5	2	40%	3	60%
			6	2	33%	4	67%
		STAT3	2	0	0%	2	100%
			15	3	20%	12	80%
			3	0	0%	3	100%
			9	3	33%	6	67%
			7	3	43%	4	57%
			9	3	33%	6	67%
			7	3	43%	4	57%
			20	7	35%	13	65%
			13	4	31%	9	69%

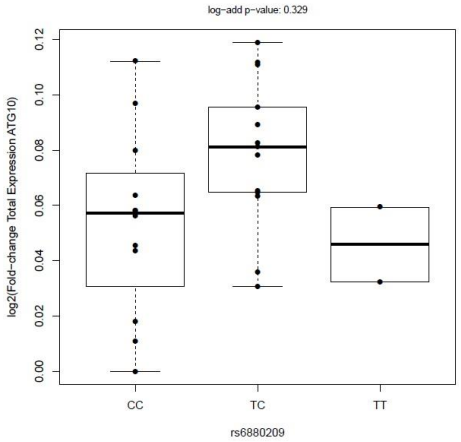
Annex 6: Total Expression of *ATG10, RPS23* and *ATP6AP1L* genes for candidate rSNPs.

The x-axis represent the candidate rSNP genotypes meanwhile y-axis represent log 2 (fold-change total expression for *ATG10* gene.

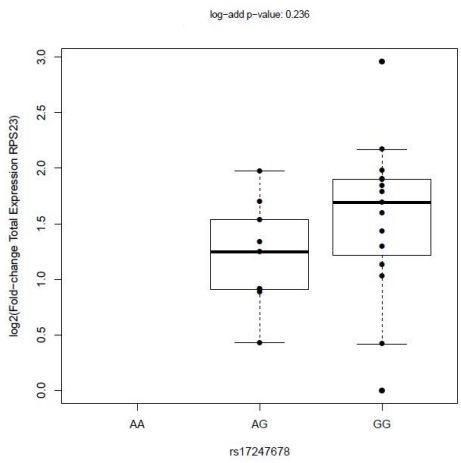


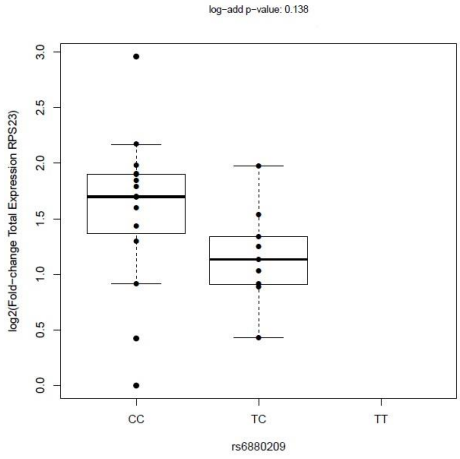


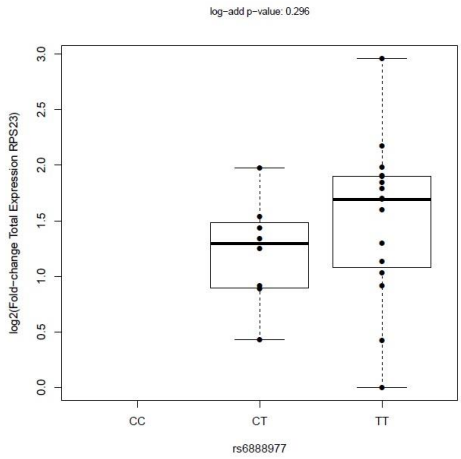


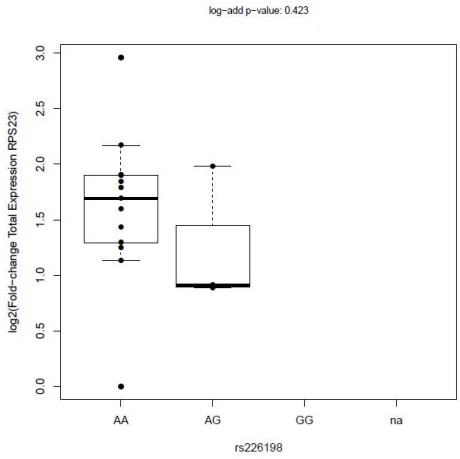


The x-axis represent the candidate rSNP genotypes meanwhile y-axis represent log 2 (fold-change total expression for *RPS23* gene.









The x-axis represent the candidate rSNP genotypes meanwhile y-axis represent log 2 (fold-change total expression for *ATP6AP1L* gene.